Some experiences teaching factorial design in introductory statistics courses

RON S. KENETT[1] **& DAVID M. STEINBERG**[2], [1]*Tadiran, Telecommunications Division, Petah Tiqwa and Tel-Aviv University, School of Engineering, and* [2]*Department of Statistics, Tel-Aviv University*

SUMMARY *A major challenge in teaching statistics is to convey how statistics is used in solving real problems. Introductory courses often fail to provide students with a sense of how to apply statistical ideas. This paper will describe our experiences with a teaching approach in which students generate and analyse their own data by conducting a factorial experiment. We note our expectations, successes, and disappointments, as well as our recommendations.*

1 Introduction

Students often complain that introductory statistics courses are a dry encounter with pages of formulas. They usually fail to see how statistics can help solve problems and to appreciate the role that statistics plays in scientific inquiry. To remedy this situation, some recent textbooks have taken a 'case study' approach (Freedman, Pisani & Purves, 1978; Moore, 1979; Cox & Snell, 1981). A strategy that we have employed is to have students run and analyse a factorial experiment as a homework assignment. In this paper we discuss our experiences, noting our expectations, successes, and disappointments, as well as our recommendations.

There are both pedagogical and statistical benefits in having students run and analyse an experiment. Most important is the hope that running an experiment will take statistics off the blackboard and into the students' daily lives. A considerable body of educational research testifies to the benefits of active engagement in learning. (The classic in this area is Bruner, 1960; see also the National Council of Teachers of Mathematics Agenda for Action, 1980). Most students find the opportunity to plan an investigation and collect their own data both enjoyable and exciting (comments rarely voiced about statistics courses!).

Despite the enormous benefits of a well-planned experiment, introductory statistics courses typically devote far more time to methods of analysis than to principles of data

collection. Two-level factorial designs are powerful data collection tools that are simple enough to be taught to beginning statistics students. Applications of statistical inference to the experiments give students a sense of how procedures such as confidence intervals and hypothesis tests are used, and how their interpretation depends on the context in which the data were collected (a connection that is too often overlooked in introductory courses). Such discussions are useful in illustrating the important link between statistics and scientific inquiry (see Box, 1976; Hahn, 1984).

Our expectation, then, was that the experimental design assignment would make our courses more interesting, more relevant, and more fun, all the while teaching a useful statistical topic and providing a unique opportunity to relate methods of analysis and data collection. We were encouraged by reports of others who have given similar assignments (Hunter, 1977 and references therein). The following sections relate some of our successes and disappointments. We conclude with a discussion of 'how to boil water', summarising data from our most popular experimental topic.

2 Description of the courses taught

Factorial experiments were assigned to several classes with students of varying backgrounds and prior knowledge of statistics: an introductory course for engineering students, mostly advanced undergraduates; an introductory course for first year social science students; a course in industrial statistics for undergraduate statistics majors; and a course in statistical quality control for graduate students in industrial engineering. The first of these courses was taught at a US university and the latter three were taught at Israeli universities. In all of these courses, the experimental design project constituted 10%–20% of the student's final course grade. Class size ranged from 15 to 40 students.

The major topic covered in the units on experimental design was two-level factorials, roughly following the format of Chapter 10 of Box, Hunter & Hunter (1978). Although the particular choice of topics differed from class to class, all emphasised graphical display, the estimation of factor effects and interactions, and interpretation of the results of the experiment. The students were taught one or more methods of obtaining an estimate of error and how to then construct confidence intervals and test hypotheses about the factor effects. Other topics included in some of the courses were: Yates' algorithm, normal and half-normal plots, interaction plots, residual plots, and reference distribution for estimated effects.

The students chose their own experimental topics and the variety of experiments was impressive. Some of the response variables studied were: time to boil water, time for a boiling egg to explode, time for a dropped ball to come to rest, time for yeast to ferment, sewing time, fraction of popcorn kernels that popped, drying time for a facial masque, time to swim 25 metres.

We briefly describe next some of the more interesting experiments conducted by our students.

1. Exposure time in developing microfiche film as a function of rinse water temperature, film speed, amount of clearing liquid, and amount of fixer. The factors acted independently and the experiment clearly indicated which had large effects and how they should be set to minimise developing time.

2. Rebound height of a tennis ball as a function of old/new, wet/dry, surface (asphalt/grass), and brand of the ball. The first three factors had large effects, but also substantial interactions. This experiment was replicated and the estimated error was quite small, so that numerous effects were statistically significant.

3. Successful sale as a function of salesperson offering service, smiling, and offering to personally find the item desired. Customers entering the store where this student worked were randomly assigned to the different test conditions and the percentage actually making a purchase computed. Each of the factors had a clear positive effect on the probability of making a purchase.

3 Difficulties encountered and recommendations

Having described above some of our 'success' stories, we now discuss some ways in which the assignment did not live up to our expectations, and some of the problems the students encountered.

The first problem with which the students must grapple is the choice of an experiment. We had hoped that they would use the assignment to study topics about which they were naturally curious, and that most would have a ready store of topics that would be appropriate candidates for an experiment. Many students found excellent topics and others used one of the topics suggested by the instructor. Several students, however, chose topics that were so silly that they obviously held no interest. For these students, evidently, this was 'just another assignment' and the data collection more of a nuisance than an adventure. The most interesting experiments tended to come from the engineering students, while social science and mathematics majors were less successful in finding good topics. Perhaps there is some relation here with the general scientific orientation of the students.

The most difficult aspect of the assignment proved to be interpreting the experimental results, in particular explaining interactions. Many reports began with a summary of the main effects that completely ignored interactions, even when some of them had been found to be quite large. We suspect that these students did not appreciate that the interpretation of interactions is directly linked to main effects, and that they should be discussed jointly. Many of these students seemed to treat interactions as 'second line' main effects, rather than as indicators of the consistency of the main effects. Perhaps this misinterpretation is reinforced by the similar procedure for calculating main effects and interactions. Other students performed detailed computations of each factor's effect at various levels of the other factors when no interactions had been detected. It is clear, in retrospect, that several examples with, and without, interactions must be discussed in class if students are to understand the meaning of an interaction.

A related difficulty in interpreting interactions concerns the distinction between 'statistical significance' and 'practical significance'. In a number of experiments, the replication error was quite small, so that many interactions exceeded two standard errors. Most students immediately concluded that these interactions were important without bothering to compare the magnitude of the interactions to the main effects. Thus, for example, an interaction whose magnitude was 5% as large as the main effects might be summarised in the same terms as an interaction estimated at 50% of the main effects, if both had the same t-statistics.

Another problem in interpretation was a tendency to phrase conclusions in generic statistical terms rather than in the substantive terms of the experiment. For example, a student who found that adding extra sugar caused yeast to ferment 2.5 minutes faster might report that 'the effect of A was -2.5'; no mention of yeast, sugar, or time. Some students failed to report the levels of the factors or to state which was high and which was low. In two experiments, it was clear that the levels of a factor were labeled incorrectly—had the students re-phrased the conclusions in terms of the experimental

variables, the errors would have been obvious. Reporting the results of an experiment may appear to be a trivial task by comparison with the problems of design, execution, and analysis, but our experiences suggest that it is an area of difficulty and should be discussed in class.

Many of the students found substantial two factor and even three factor interactions in their experiments. We suspect that one of the reasons for this was that the response variable was often measured in time. Time is a convenient measurement scale for home experiments because the only equipment required is a watch. Unfortunately, many phenomena are not additive on a time scale, so that interactions are found. Often a simple transformation of the response variable will reduce the extent of interactions (for a classic example, see Box & Cox, 1964), so experiments involving time can provide an excellent springboard for introducing transformations. If, however, discussion of transformations exceeds the scope of the course, we suggest that the instructor describe, as examples, a number of experiments in which the response is not measured in time.

A number of students did not understand the basic format of conducting an experiment. Some of the statistics majors, who had some past experience analysing data, asked the instructor if they could use data from an existing medical data bank. One of their proposed explanatory variables was continuous, but they planned to dichotomise it in order to perform the standard factorial analyses. Other students also reported explanatory variables that were either not controlled or dichotomised versions of continuous variates.

One area of major disappointment was 'phony' data. One student discussed an experiment on avocados in which the response times averaged roughly 100 days—quite remarkable for a three week assignment! Even more astounding was the discovery that the experimental results were reproducible—a classmate measuring flower life (in hours) obtained absolutely identical data. Although this was the only blatant case of fabrication, our personal probabilities regarding several other experiments are not small. We believe that actually conducting the experiment is a crucial part of the assignment. It is important that students realise the difficulty in setting up an experiment, in fixing levels of explanatory variables and measuring response variables, in randomising run order, in the possible need for blocking, etc. These concepts often take on new life when they are associated with real problems in experimentation.

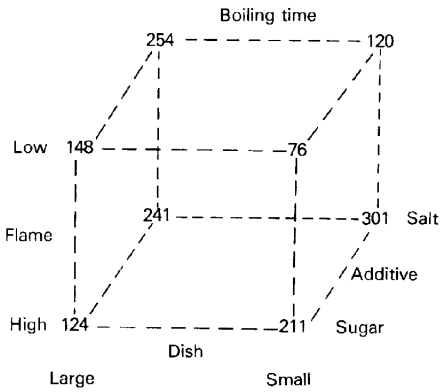

FIG. 1. Plot of the data from Experiment 1. Boiling time is reported in seconds.

4 The boiling water experiment

The most popular experimental topic was to study the time required to boil water. In this section we briefly describe five such experiments. The figures and occasional reanalyses were done using the JASS^R software package (1985) for factorial designs.

Experiment 1

The data in Fig. 1 come from an experiment in which the factors were flame (high/low), dish size (large/small), and additive (sugar/salt), with time to boil measured in seconds. The data, as reported, are quite remarkable: the water boiled faster with a low flame than with a high flame! The inescapable conclusion is that the data have been reported incorrectly. (Perhaps the student randomised the run order but then filled in the results as though the standard order had been used.) These data illustrate the importance of expressing experimental results in terms of the actual factors, rather than referring generically to 'the effect of factor A'. Had the student done so, the mistake would no doubt have been discovered.

TABLE 1. Estimated factor effects, standard errors, and t-ratios for Experiment 2

	Effect	Standard error effect	t-ratio
Average	164.656	0.471	349.716
Flame (F)	−39.438	0.942	−41.881
Cover (C)	−11.313	0.942	−12.013
Sugar (S)	9.563	0.942	10.155
FC	−0.563	0.942	−0.597
FS	−1.188	0.942	−1.261
CS	1.188	0.942	1.261
FCS	−0.063	0.942	−0.066

S=2.663 with 24 d.f.

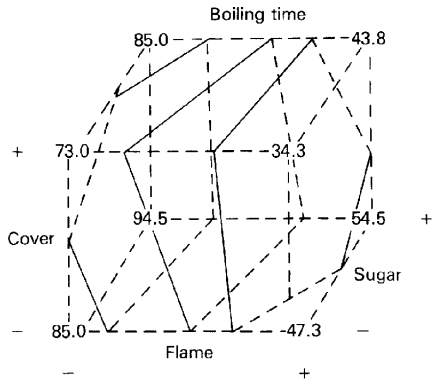

FIG. 2. Plot of the data from Experiment 2. Boiling time is reported in seconds and each number has been reduced by 100 seconds. The added lines display surfaces of equal response.

TABLE 2. The data reported for Experiment 3, a replicated 2^3 on the time required to boil water

Factor settings			Boiling time (minutes)	
Flame	Cover	Dish	Rep. 1	Rep. 2
Low	No	Small	10.50	10.65
High	No	Small	2.01	1.59
Low	Yes	Small	4.16	4.14
High	Yes	Small	1.58	1.57
Low	No	Large	5.26	5.27
High	No	Large	1.58	1.54
Low	Yes	Large	4.10	4.05
High	Yes	Large	1.54	1.48

Experiment 2

In most of the experiments on boiling time, large interactions were discovered—the data in Fig. 2 are an exception to this rule. The factor effects and standard errors listed in Table 1 show that each of the three factors had a strong and consistent effect on boiling time. The interactions are small and, even with the large number of replicates (four at each factor combination), have small t-ratios as well. We added (manually) lines representing surfaces of equal boiling time to Fig. 2, an effective way to display the conclusions reached from Table 1.

Experiment 3

This experiment illustrates two of the pitfalls of measuring the response in time: ambiguous recording and the need for a transformation. Table 2 lists the data reported by the student: two replicates were run at each corner of the cube. Experimental error was quite small, except for the measurements 2.01 and 1.59 in the second row of the table. We suspect that these measurements are, in fact, 1 minute 59 seconds and 2 minutes 1 second. The 10.65 reading in the first row seems to indicate that our hunch is wrong; after row one, however, none of the measurements have a decimal part larger than 0.59, which suggests that the method of recording may have been altered in mid-experiment! Unfortunately, we have no way of knowing whether or not our hunch is correct.

Accepting the data as reported, Table 3 shows the analysis of boil time. Although the size of the flame is the dominant effect, all the interactions are large and no simple explanation is possible. When the data are transformed from time to speed (by taking inverses), a much different picture emerges (see Table 4). The interactions all but disappear (only the cover by dish interaction is noteworthy) and the dominance of the flame effect is much more clearly established (it is 6 times as large as the next largest effect on the speed scale, but only 2.3 times as large on the time scale). The four direct comparisons of high flame to low flame are much more consistent: the increases in speed range from 0.394 to 0.468 where the corresponding decreases in time ranged from 2.5 to 8.8.

TABLE 3. Estimated factor effects, standard errors, and *t*-ratios for Experiment 3, analysed on the time scale

	Effect	Standard error effect	*t*-ratio
Average	3.8138	0.0284	134.0530
Flame (F)	−4.4050	0.0569	−77.4179
Cover (C)	−1.9725	0.0569	−34.6667
Dish (D)	−1.4225	0.0569	−25.0004
FC	1.8350	0.0569	32.2501
FD	1.2700	0.0569	22.3202
CD	1.3525	0.0569	23.7702
FCD	−1.2650	0.0569	−22.2324

S=0.1138 with 8 d.f.

TABLE 4. Estimated factor effects, standard errors, and *t*-ratios for Experiment 3, analysed on the speed scale

	Effect	Standard error effect	*t*-ratio
Average	0.4091	0.0084	48.4340
Flame (F)	0.4327	0.0169	25.6168
Cover (C)	0.0737	0.0169	4.3653
Dish (D)	0.0513	0.0169	3.0384
FC	−0.0272	0.0169	−1.6100
FD	0.0014	0.0169	0.0842
CD	−0.0353	0.0169	−2.0904
FCD	0.0102	0.0169	0.6009

S=0.0338 with 8 d.f.

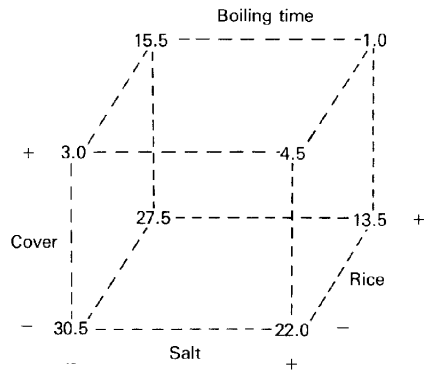

FIG. 3. Plot of the data from Experiment 4. Boiling time is measured in seconds and each number has been reduced by 100 seconds.

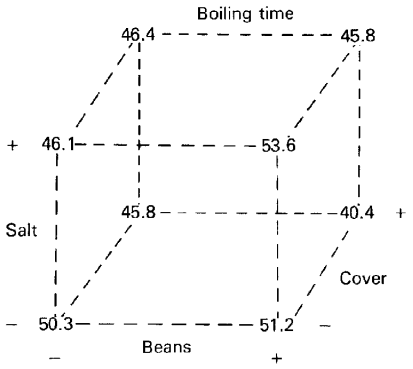

FIG. 4. Plot of the data from Experiment 5. Boiling time is measured in seconds and each number has been reduced by 100 seconds.

Experiments 4 and 5

These experiments used similar factors, but reached slightly different conclusions. In both experiments, the factors were cover (yes/no), salt (yes/no), and presence of a substance to be cooked (rice in experiment 4, beans in experiment 5); the data are shown in Figs. 3 and 4, respectively. The dominant factor was the cover, which reduced boiling time on average by 17 seconds in experiment 4 and 6 seconds in experiment 5, but the effect of the cover was not altogether consistent. The major

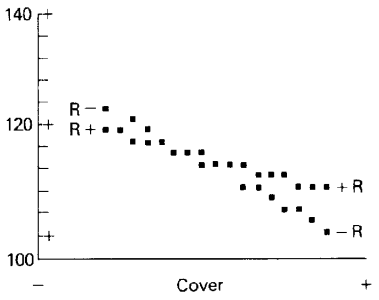

FIG. 5. The cover by rice interaction plot for Experiment 4.

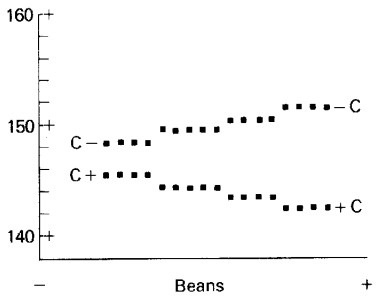

FIG. 6. The cover by beans interaction plot for Experiment 5.

factor interacting with cover was the cooking substance, but its effect in the two experiments was opposite! The interaction plot in Fig. 5 shows that the effect of the factor interacting with cover was the cooking substance, but its effect in the two experiments was opposite! The interaction plot in Fig. 5 shows that the effect of the cover was more pronounced when no rice was in the water (22 seconds versus 12 seconds with rice), but the plot in Fig. 6 shows that, in experiment 5, the cover had almost no effect without beans (2 seconds) but had a strong effect when the beans were added (9 seconds).

Finally, it is interesting to note the extreme contrast in precision of the two experiments. The estimated standard errors in experiment 4 are almost 15 times larger than those in experiment 5.

Correspondence: Ron Kenett, Tadiran, Telecommunications Division, P.O. Box 500, Petah Tiqwa, Israel.

REFERENCES

BOX, G.E.P. (1976) Science and statistics, *Journal of the American Statistical Association*, 71, pp. 791–799.

BOX, G.E.P. & COX, D.R. (1964) An analysis of transformations (with discussion), *Journal of the Royal Statistical Society Series B*, 26, pp. 211–252.

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

BRUNER, J.S. (1960) *The Process of Education* (Cambridge, MA, Harvard University Press).

COX, D.R. & SNELL, E.J. (1981) *Applied Statistics: principles and examples* (New York, Chapman & Hall).

HAHN, G.J. (1984) Experimental design in the complex world, *Technometrics*, 26, pp. 19–31.

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

JASS^R *Reference Manual* (1985) (Madison, WI, Joiner).

MOORE, D.S. (1979) *Statistics: concepts and controversies* (San Francisco, W. H. Freeman).

NATIONAL COUNCIL OF TEACHERS OF MATHEMATICS (1980) *An Agenda for Action: recommendations for school mathematics of the 1980s* (Reston, VA, National Council of Teachers of Mathematics).