TECHNIQUES USEFUL IN TEACHING A FIRST COURSE ON STATISTICAL DESIGN

by

Walter T. Federer

Abstract

Teaching statistical design to undergraduate and graduate students inexperienced in experimentation, requires special attention. Experimentation on a computer is not the answer and neither is obtaining statistical design by definition as is so frequently done in statistical methods and mathematical statistics textbooks. Planning and conducting real-life investigations should be a part of a first course on design. Also, class demonstrations and well-selected examples and problems are very helpful. Some of the above, which were used in a second year undergraduate course, are described.

TECHNIQUES USEFUL IN TEACHING A FIRST COURSE ON STATISTICAL DESIGN

by

BU-817-M May, 1983

Walter T. Federer

Introduction

When an undergraduate and a first year graduate student first encounters the topic of statistical design, their comprehension and appreciation depends upon the amount of research or other investigational experience they possess, how the topic of statistical design is presented, the examples, problems, and lecture demonstrations. The more immature the student, the greater the necessity for paying attention to one's approach. Also, since most statisticians do not conduct investigations and experiments on anything but a computer, they are somewhat ill-equipped to teach a course on statistical design. To the majority of statisticians, as judged by their writing and comments, the term "experimental designs" means combinatorics of constructing plans, development of associated linear model theory, statistical methodology, computing and computer programming, significance testing, and/or hypothesis testing for data from something called an "experimental design". The planning of an investigation is foreign to this group. The population from which the experiment is a sample, the nature of sampling units, experimental units, observational units, replication, and reasons for using certain error terms are not discussed except in vague and fuzzy terms. This vagueness and fuzziness is passed on to students; it is small wonder then that many experiments, surveys, and other investigations are poorly planned and explained. It is very difficult to find an article in a scientific journal which clearly and precisely describes the manner in which the investigation was conducted.

Paper No. BU-817-M in the Biometrics Unit, Cornell University, Ithaca, NY 14853.

Now Statistics is composed of three major areas, viz.

(i) planning an investigation (statistical design),

(ii) summarizing the results of an investigation (statistical methods and computing procedures), and

(iii) making inferences from the investigation to the population to which it pertains.

Since the main emphasis of Statistics over the past 50 years has been on (ii) and (iii), one can easily understand the present status of Statistics. Computer program packages make the situation worse as no thought is given to the nature of the data. Some administrators feel that a statistician can be replaced by a computer package. They may be correct for certain types of statisticians. Today's and tomorrow's investigations deserve better statistical treatment than they are receiving. Statistical design, or the commonly used term "experimental design", in the real world of experimentation is something quite different and much more inclusive than solely combinatorics, response surface methodology, sample size determination, linear model theory, statistical computing, etc. Statistical design consists of at least the following entities:

(i) Precise description of population structure for each variable under study is necessary. Precise definitions are required for the sampling units (elements making up the population), the experimental unit (the smallest unit to which one treatment is applied), the observational unit (the smallest unit on which observations are made), and the response model equations both before and after the treatments are applied.

(ii) Measurements and measuring instruments must be described and necessary calibrations made. This topic is not mentioned in statistics textbooks.

(iii) Survey design consists of procedures for selecting a sample of sampling units from the population.

(iv) Treatment design refers to the selection of treatments for an experiment, including checks, standards, and placebos.

(v) Experiment design consists of the plan, layout, or arrangement of treatments in the experiment.

(vi) Modeling and model design consists of determining an appropriate and/or a correct response model. It does not assume a linear model structure. Model design is the plan of obtaining experimental units for determining a correct, or at least an appropriate, response model.

(vii) Sequential design refers to the method of taking observations sequentially in an experiment where the previous results are used to determine which observations are taken next.

(viii) Sample size and allocation relates to the number of experimental units, or number of replicates, to be taken and how they are to be distributed over the strata in the population.

(ix) Techniques for conducting investigations relate to accuracy checks, control of bias, to size and shape of sampling or experimental units, and to independence of results in the experiment including control of competition.

(x) Principles and properties of statistical designs are many and varied. They are needed to demonstrate to students the relative value of two statistical designs or classes of designs.

Some examples, demonstrations, problems, and projects are described below. These have been found useful in teaching a course to undergraduates without research experience. The course is entitled "Statistics and the World We Live In" and the text is Statistics and Society written by the author (1973). The descriptions are given mostly in the order in which they appear in the course.

Problem 1 (first problem assigned):

A table is of unknown length and unspecified shape. The measuring instrument is calibrated in feet. You are not allowed to calibrate the instrument; this might be the logical thing to do in some circumstances but in others it might be impossible. Give a procedure for taking a large number of measurements which would allow accuracy up to 1/100 of a foot, say.

The purpose of this problem is to demonstrate that ingenuity in design is necessary for experimentation. There are several possible procedures for this, five of which are described in the "Supplement to Statistics and Society" (Federer, 1972). It also emphasizes the nature of measuring instruments and measurements.

Discussion (week 2):

In order to acquaint the students with the nature of measuring instruments and measurements, the following measurements are taken by two groups of individuals in the Discussion period:

Group A	Group B
height measured in yards	eye color
" " " feet	hair color
" " " millimeters	height measured in millimeters
weight " " pounds	

In addition, students give their own weight in pounds, height in inches, hair color, eye color, age, and class standing. The measuring instrument for Group A is biased a positive three centimeters but students are not told this. Also, the scale is set at a negative five pounds to show need for checking instruments.

Additional objectives are to show students that although a census was planned, this was not achieved, that no variation in height in yards results, that only two classes, 5 feet and 6 feet, results when height is measured in feet, that no two people have the same height in millimeters when measured by two different people, that some people refuse to have their weight taken, that a large mass of data is quickly accumulated, and that one needs summary procedures for eliciting the facts from this survey. The data are tabulated in tables and distributed to the students when summarization of data is discussed (Chapter 8, Federer, 1973). These data have a personal meaning for the students as they came from the class members and were taken by class members.

Lecture - phony statistics:

Phony statistics are amply demonstrated. Examples abound in the writings of Campbell (1974), Huff (1954), Seligman (1963), and Wilson (1952). One note of caution, although it is easy to show how to obtain bad data and conclusions, this is a negative approach which must be balanced with a positive approach. The phony statistics examples are used to demonstrate the need for following the steps in scientific experimentation (Wilson, 1952), survey design, experimental design, and treatment design (Chapters 4, 5, 6, and 7 of Federer, 1973).

Discussion (week 5):

During the discussion period of the fifth week of class and just before the class members are being urged to start their class projects, a questionnaire of seven questions is given to all members of a discussion group. The seven questions and the number corresponding to an answer are:

1. Are you under 21 years of age?

   Yes (0) No (1)

2. Did you cheat in any way on the Stat 200 prelim that you took last week?

   Yes (2) No (3)

3. In general, are you happy with your decision to come to Cornell?

   Yes (1) No (0)

4. While at Cornell, have you ever stolen money or any other article worth over $5.00 from a roommate, friend, employer, or anyone else?

   Yes (3) No (2)

5. Does your parent earn more than $25,000 a year?

   Yes (0) No (1)

6. Have you smoked any marijuana during the past two weeks?

   Yes (2) No (3)

7. Are you enrolled in the College of Agriculture and Life Sciences?

   Yes (1) No (0)

Questions 2 and 4 are sensitive questions which students will not answer correctly if a direct method of questioning is used. Four methods of obtaining answers to these questions are:

(i) randomized response for three of seven questions using three randomizing devices and one of the following sets of three questions for one student:

(1,2,4), (2,3,5), (3,4,6), (4,5,7), (5,6,1), (6,7,2), and (7,1,3).

These were the seven sets of three obtained from a symmetrical balanced incomplete block design.

(ii) block total response — A student had one of the above seven sets of three questions and he was asked to give a total for the three questions using the numbers indicated for a question.

(iii) randomized form of block total response — A student randomly selected one of the seven sets of three questions and gave the block total. His selection was anonymous. Not only did the student not give individual answers but the interviewer did not even know which set of three questions was being answered.

(iv) supplemented block total — Here the sets were questions 1 and 4, 2 and 4, 3 and 4, 5 and 4, 6 and 4, 7 and 4, and all seven. A student gave only a total for two questions or of seven questions, depending upon the set being answered. Question 4 was the most sensitive question.

The four methods of obtaining anonymous responses were presented to students for possible use in their class projects as well as to acquaint them with techniques which could give reliable answers to sensitive, incriminating or embarrassing questions. Several students used one of the methods to obtain information on sex-related questions in their class projects.

Demonstration - Weighing design:

To demonstrate that one experimental procedure can be more efficient than another, consider the following four methods for weighing seven objects on a balance scale in eight weighings:

(i) One weighing is made to balance the scale and each of the seven objects is weighed once to produce eight weighings.

(ii) All objects are weighed and then seven groups of three objects are weighed. The groups are:

(1,2,4), (2,3,5), (3,4,6), (4,5,7), (5,6,1), (6,7,2), and (7,1,3).

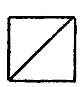
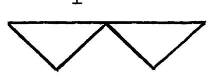
(iii) All objects are weighed and then seven groups of four objects are weighed. The groups are:

(3,5,6,7), (4,6,7,1), (5,7,1,2), (6,1,2,3), (7,2,3,4), (1,3,4,5), (2,4,5,6).

(iv) All objects are weighed and then seven groups of six objects are weighed. The groups are:
(1,2,3,4,5,6), (2,3,4,5,6,7), (3,4,5,6,7,1), (4,5,6,7,1,2),
(5,6,7,1,2,3), (6,7,1,2,3,4), (7,1,2,3,4,5).

If the variance of a single weighing is σ^2, the variances of weights of any object for (ii) is $\sigma^2/2$, for (iii) is $0.47\sigma^2$, and for (iv) is $43\sigma^2/49$. Thus if one uses either (ii) or (iii), one can cut the variance in half over (i). In fact, (iii) is only $0.47\sigma^2$, indicating that it is the optimal design in this class. Weighing two or five objects at a time would not give a balanced design and would give a higher average variance than (ii) or (iii).

Demonstration - Randomized complete block design:

This problem is presented to demonstrate the actual layout of an experiment designed as a randomized complete block design. Cardboard blocks 4" x 4" x 24" are constructed. One could use actual wood blocks but they are heavy. It is supposed that one has a large pile (population) of wood blocks 4" x 4" x 24" and that one wants to compare four brands of paint for their durability to weathering and wood preserving qualities over a period of five years. A simple random sample of blocks is obtained. The minimum size experimental unit (e.u.) allowable is an area 4" x 12" for a paint. The first problem is to decide on an e.u. One could use one side for each of the 4 paints, or one could use a 6" strip around the block. To make the e.u.'s as alike as possible one could cut the block diagonally into two pieces as  to give the two halves as $\sqrt{32}$" x 24" with an e.u. being $\sqrt{32}$" x 12". This would give four e.u.'s as alike as possible; they are only the width of a saw blade apart. The two halves could be painted with the four paints and laid face-up  for uniform weathering. If one side of a block was an e.u., the block would need to be rotated in order to allow equal exposure to sun and shade. The four paints are randomly allotted to the four e.u.'s and r blocks are used.

This example demonstrates the selection of e.u.'s and the handling of e.u.'s during the course of an experiment. It is visual and easily comprehensible. It can be demonstrated that there is more block to block variability than there is among e.u.'s within a block.

Examination question - Balanced incomplete block design:

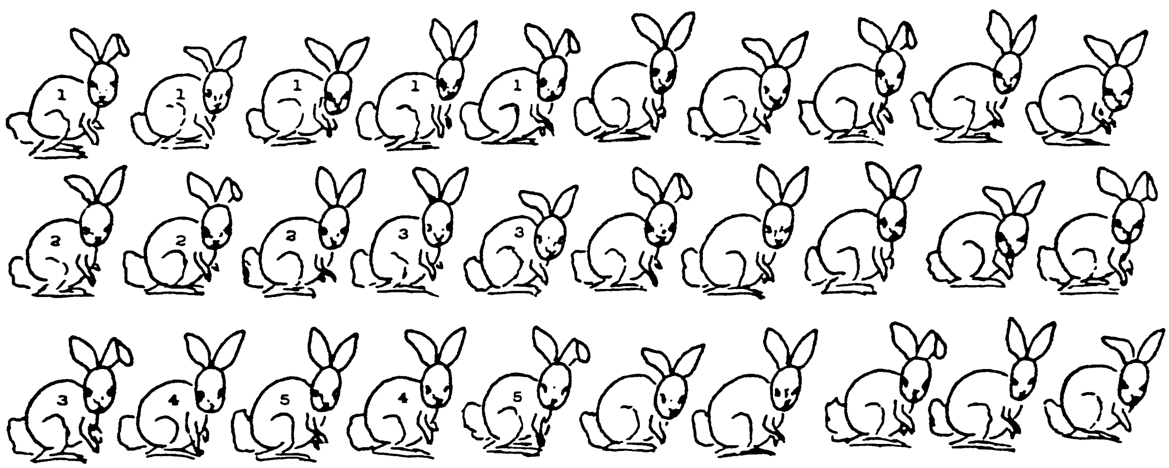
A sample prelim (examination) for each of three prelims is given on pages 360-379 of Federer (1973). One of these questions is:

3. Mr. Bunny Buck wishes to compare v=5 feeding treatments on the growth of rabbits from three to six weeks of age. Since he has available to him only litters of 3 rabbits each he decides to use a balanced incomplete block design.

   Construct a balanced incomplete block design for v=5 treatments in blocks of size k=3 rabbits each by completing the following:

Litter number = Block number

1 2 3 4 5 6 7 8 9 10



The number of replicates is r = ____________________.

The number of times any pair of treatments occur together in the b incomplete blocks is λ = ____________________.

Which of the assignable effects are orthogonal to each other?

Describe and illustrate the randomization procedure for this design using the attached random number table starting at the top of column 3 and proceeding down the column. (Write down the sequence of random numbers used.)

A set of random numbers was provided.

This example gives a visual impression of how a balanced incomplete block design could be used when there were more treatments, v = 5, than could be accommodated on k = 3 male rabbits in each litter (block). The answers to the test questions are given in Federer (1972).

Demonstration - Latin square and orthogonal Latin square designs:

The late W. J. Youden used the example described in Federer (1973), pages 90-93, to illustrate the use of experiment design techniques to control various sources of variation and how to design the treatments, brands of oil, to be orthogonal to the other sources of variation. This example demonstrates blocking on four sources of variation. In the demonstration, five empty oil cans are used as well as four model cars. These can all be placed in a box and pulled out as the story (see Federer, 1973, pages 90-93) goes along. Overheads and overlays can be used effectively along with this demonstration.

Examples:

Various examples encountered over the years cause considerable student interest. This is actual interest rather than instructor-presumed interest in an example. The first example illustrates the need for careful treatment design in planning investigations and the need for adequate controls. The data were obtained by Jellinek, 1946, Biometrics 2, 87-91, and are presented and discussed on pages 143-148 in Federer (1973). Two types of checks, the standard drug treatment and a placebo, were needed, the latter to distinguish between individuals with psychological headaches and with physiological headaches. The standard drug was needed as a check to determine the effectiveness of the other two drugs.

The second example was the results of 26,306 throws of 12 dice by Weldon, as described in Fisher (1925), coupled together with the frequency distribution of number of boys per family in 53,680 families of size eight. These data for the latter were obtained by Geissler and are given in Fisher (1925) as well as in Federer (1973). The dice example demonstrated a poor fit to a binomial distribution with p = 1/3 but an excellent fit to the binomial for p obtained from the biased data. Geissler's data do not fit a binomial distribution although one could think of several reasons why they should. Students speculate about the lack of fit.

The third example was Bortkewitch's data for the number of deaths per year per army corps from horsekicks (see Federer, 1973, page 299). These data give an excellent fit to a Poisson distribution. Together with the above example, one needs to have a set of data which might reasonably be expected to be Poisson but is definitely not. The set of data on number of decayed teeth in boys from Rao and Chakravarti (1956), Biometrics 12, 264-282 (see Federer, 1973, page 300), is of this nature. Examples of the above type have provoked students to think of reasons why a set of data does or does not conform to a particular distribution. An example which students consider "gross" is the number of accidents over a period of five weeks for 647 women working on high explosives in a factory in England during World War II. The data give a good fit to the negative binomial. Students never inquire about the nature and frequency of accidents for this example. It does not appear to make them curious, just "grossed out".

Class Project:

The single most important and well-liked part of the course "Statistics and the World We Live In" was a class project. Students working in self-selected groups of one to five, but usually two or three, selected a topic on which they wished to plan an investigation. They were to proceed following the steps of scientific investigation outlined in Chapter IV of Federer (1973). In their written report they were to detail how these steps were followed. The project had a grade value of two prelims and was given twice the value of the fourteen weekly problem sets. Since it constituted one-third of the grade before the final examination, considerable effort was expected and given. The groups were free to select the type of investigation which could be a survey, an experiment, an observational study, or a study of a procedure. The project topics were many and varied. A list of these is given each year in the Annual Report of the Biometrics Unit. Several projects each year cause changes in one or more aspects of campus life and activities, several have been published in technical journals, and a few have had impact in politics. The importance of Statistics becomes especially well demonstrated to these students.

Another method for obtaining experience in the actual conduct of experiments would be to design and conduct a series of investigations planned by class members and supervised by the instructor. For example, one could compare the heights of bounce for five brands of golf balls dropped from several different heights. Several replications could be used with different balls each time.

One might think that M.S. and Ph.D. students in Statistics would have little or no difficulty in describing precisely how to obtain a randomized plan for a randomized complete block design, a balanced incomplete block design, and a Latin square design. To demonstrate that this is a problem, it is suggested that the reader's next ten graduate students be asked to do this on an oral, or even a written, examination. Another question is to have them describe precisely the sampling unit, the experimental unit, and the observational unit for any selected experiment. The reader will be amazed how poorly they do on problems of this nature. If they cannot answer such questions as the above in clear and precise terms, how can they possibly teach students to do this? They may be able to demonstrate very clearly and precisely that a statistic is asymptotically normal or that an experiment design is asymptotically orthogonal and/or balanced, but they most likely would be unable to explain what all this means to an experimenter who cannot go to the limit but who must remain in his finite world of experimentation.

Literature Cited

Campbell, S. K. (1974). Flaws and Fallacies In Statistical Thinking. Prentice-Hall, Inc., Englewood Cliffs, New Jersey. pp. 200.

Federer, W. T. (1972). Supplement: Statistics and Society. Biometrics Unit, Cornell University. pp. 155.

Federer, W. T. (1973). Statistics and Society: Data Collection and Interpretation. Marcel Dekker, New York. pp. 399.

Fisher, R. A. (1925). Statistical Methods for Research Workers (1st of several editions). Oliver and Boyd Ltd., Edinburgh. pp. 350.

Huff, D. (1954). How to Lie With Statistics. W. W. Norton and Company, Inc., New York. pp. 142.

Seligman, D. (1963). We're drowning in phony statistics. In Statistics in Action (editor: T. J. Sielaff). The Lansford Press, San Jose, California. pp. 251.

Wilson, E. B. (1952). An Introduction to Scientific Research. McGraw-Hill Book Company, Inc., New York. pp. 373.