THE ANATOMY OF A STUDY[1]

N. Scott Urquhart[2]

Department of Experimental Statistics, New Mexico State University, Las Cruces, NM 88003

Formally, anatomy deals with both the actual and conceptual isolation of various parts and systems of an organism for the purpose of describing its parts, their positions, relations, structures, and functions. Although you may associate memorization with anatomy, it really serves to structure our thinking about an organism. In fact, the human mind seems to understand a complex situation or structure by first subdividing it into components, then studying each component and finally relating the components. Reflect for a minute: we do this as teachers, researchers, and students. It seems reasonable, even though perhaps not formally correct, to describe the dissection of a study into its parts, their examination, and their relations as the anatomy of a study. I plan for the following anatomy to focus your attention on general structures without particular concern for occasional difficulties and degeneracies.

A research study

A *research study* usually seeks to describe how a particular response changes with certain related features of the research setting. Studies must be done in the face of natural variability. If you understand some determinants of this variability, you may minimize its impact by how you structure your research study. This brief introduction isolates the 3 essential parts of a research study: 1) the response, 2) the structure to which we want to relate the response, and 3) the structure through which we seek to minimize the impact of extraneous variability. I choose to examine each of these parts under the titles of the *response design,* the *treatment design,* and the *experimental design.* The names for these parts will assume a more intuitive meaning as we progress.

A RESEARCH STUDY HAS A

- *Treatment Design,*
- *Experimental Design, and*
- *Response Design*

Examples will help communicate the points. We could use bits of published studies, or we could discuss the entire anatomy and illustrate it with one grand example. My teaching experience suggests a more understandable alternative: We will consider an example in some detail now, and use it for recurring illustrations.

This example really is a collage of several studies that I have worked with: It is just complex enough to clearly display a study's essential parts without becoming bogged down in distracting details. As you proceed through its anatomy, consider illustrations from your own research. Hopefully you will find them illuminating.

Study's objectives. Water quality is an important matter, particularly in arid regions. Current definitions of water quality focus principally on total dissolved solids. Some dissolved solids may be biologically irrelevant over wide ranges of concentrations, while small variations in other dissolved solids may have important biological consequences. Thus, consider comparing the quality of different water sources with a biological assay by comparing growth of an organism grown with water from the different sources.

Study's structure. Suppose chrysanthemums were used as the assay organism; they are fairly easy to work with and homogeneous cuttings are available commercially. Water was obtained from 24 sources in sufficient quantities for raising the mums and for associated chemical analyses on the water from each source. The water sources ranged from distilled water to tap water to brackish water to water from sulfur springs. The mums were grown in pots (360) in an research greenhouse. The pots were placed on 3 benches, 24 groups of 5 pots on each bench. Each water source was randomly allocated to a group of 5 pots on each bench with a separate randomization for each bench. We will consider both 1 cutting per pot and 4 per pot. Suppose that, after 7 weeks, plant height was evaluated as a biologically relevant assay variable.

This study's *treatment design* revolved around how the water sources relate to each other through their chemical composition. A restriction was placed on where the treatments (the water sources) could appear in the experimental area. Specifically, each treatment had to appear in the same number of pots on each bench. This tells us that the study was conducted in a randomized complete block experimental design, an important *experimental design,* which is familiar to most horticultural scientists. The experimental units consisted of sets of 5 pots to which a water source was applied. Each pot represents a response unit when the pots contained a single cutting. These latter considerations imply that this study has a simple hierarchical, or nested, *response design.*

If instead, each plant's height had been obtained weekly, 2520 (7 × 360) observations would have resulted, and the response design would have involved repeated evaluations through time. Or, if there had been 5 cuttings per pot, evaluated only once, the response design would have been a 2 level hierarchical response design (5 plants per pot, 5 pots per treatment).

A general statistical structure

You already have encountered the terms treatment and experimental unit. As these terms frequently will appear here, you need more than an intuitive familiarity with them. Here is a statistical structure that gives the needed meaning across a broad range of situations.

In the example, water sources functioned as treatments and sets of 5 pots were the experimental units. On the other hand if the 120 (5 × 24) pots had been randomly assigned to the 120 positions on the bench, then the pots themselves would have been the experimental units. More generally, when we set the conditions under which research material will be kept, the conditions usually are the treatments and the amount of material to which we apply a treatment is the experimental unit. Of course, an experimental unit can consist of several parts, such as several pots in a unit, several plants in a pot, or several rows in a field plot with a response coming from each part. A fuller discussion of this situation follows when considering the response design.

We may fairly describe experiments as studies in which the investigator applies treatments to research material, but many research studies do not allow the researcher this option. A researcher cannot alter the cultivar of a lettuce seed, the age of an existing tree, the type of soil in a plot, the sex or socio-economic status of students in a classroom, or the manufacturer or size of a tractor tire. These are similar to our earlier treatments, but the research material determines its treatment status by its characteristics. If the researcher cannot *apply* the treatment, then the previously stated concept of an experimental unit does not apply. Our conceptual framework must be expanded to include these "characteristic treatments" because most research studies with a rich treatment structure somehow involve them.

General structure. A *population* consists of a well-defined set of objects for which inferences are sought, like pots of chrysanthemums, apple trees in Indiana, or chili plots in New Mexico. Think of a set of populations, each population composed of experimental units and one population for each treatment. This requires a broad

[1] Journal Article 839, Agricultural Experiment Station, New Mexico State University, Las Cruces, NM 88003.
[2] Professor.

An experimental unit is an object randomly selected from the population.

e.g. — A bulb selected from a lot

or

A plot selected to receive 200 ppm of a chemical

conception of a treatment: it essentially consists of a way to index or describe the distinctive features of the populations we wish to study. The index may be a single integer or letter in simple situations; several integers or letters in more complex situations; continuous index values for regression; or some combinations of these for covariance. So, the treatments describe those features of the populations to which we hope to relate the response and whose importance we hope to evaluate through our research study.

We will use the word *treatment* as a general term for a variety of distinctive features of populations; its use neither implies nor requires that you treat some of your material differently from other parts of it. Specifically, treatments could involve strains of a pathogen, methods of storing apples, kinds of deer, chickens or pigs, processes of pasteurizing milk, sires in a dairy or beef herd, ways of vaccinating an animal; diets for plants, humans, or animals; human races, sizes of stones, amount of heat, time, or weight, etc!

Experimental units. This still leaves the idea of an experimental unit somewhat vague. Each population simply consists of the experimental units. To examine the populations statistically, we must get random samples of experimental units from each population. This usually leads us to the definition of an experimental unit in practice: an experimental unit represents the basic element we select from the populations. We may later subdivide it or measure it several times, yet its random selection defines it, regardless of how many pieces of data it yields.

Now we can return to consider the 3 basic parts of a study.

Treatment designs

The treatment design specifies the relative structure of a set of treatments; specifically, how the treatments relate to each other for purposes of examining the response. This implies experimental and, in turn, statistical questions which should be examined.

The experiment using mums to assess water quality involved 24 treatments: the water sources. If the investigator did nothing more than obtain water from 24 sources, he would have an unstructured set of treatments. If the water sources had come from 4 localities, then comparisons among groups of treatments would reflect location differences. This sort of structure is called a *grouping treatment design*. Instead, the responses could be related to a particular quantitative characteristic(s) of the water such as total dissolved solids, iron content, arsenic content, etc. Such treatment structures will be called *gradient treatment designs* and include, as a subset, situations that we commonly associate with regression. Finally, the 24 water sources could have been a random sample of all possible water sources in some area. This type of structure is called a *random treatment design* or Model 2 set of treatments.

Composite treatment designs. The 24 water sources could have more structure than simply being different. They could have resulted from 4 dilutions (X, 0.75X, 0.50X, 0.25X) of water originally drawn from 6 sources. We could naturally arrange the results from such a study in a table having 4 rows (for the dilutions) and 6 columns (for the water sources). This provides a simple example of a *factorial treatment design*. More generally, a factorial set of treatments results from all possible combinations of 2 or more basic treatments, usually called factors. Thus, we can examine the overall effect of each factor using its own treatment design and also study how the factors interact. Frequently, the factors' designs suggest fruitful ways to study this interaction.

TREATMENT DESIGNS

- *Identifies relations of central concern*
  Specifies how various populations relate to each other for the purpose of drawing subject matter conclusions
- *Requires definition of treatment populations*
  What populations are going to be studied?

Most other composite treatment designs result from either adding or discarding treatments from a complete factorial set of treatments. For example, fractional factorials and response surfaces represent 2 important specialized extensions. A fractional factorial treatment design merely contains part of a regular factorial set. Judicious choice of this part or fraction can yield interpretable results when large numbers of factors and/or large numbers of levels occur. An unplanned fractional factorial may just happen: consider what happens to a complete factorial if it loses a cell (or several) by accident or because no response is possible under certain conditions your factorial set produces. Other fractional factorials do exist and have instances of application in a wide variety of disciplines including the plant, animal, and behavioral sciences, but further discussion of them here would digress from the anatomy. See Cochran and Cox (3) for details on this topic.

Another variation of a factorial treatment design is a *response surface design*. They were developed especially to isolate optimum operating conditions in industrial situations. They are equally applicable in any science seeking to isolate optimum substrate conditions — production economics, for example. Thus, they should be appropriate for production studies in large commercial greenhouses. See Box et al. (1) for a further discussion of this topic.

KINDS OF TREATMENT DESIGNS

UNSTRUCTURED

- *Fixed—cultivars, cultural practices; use multiple comparisons*
- *Random—years, locations, genetic lines; use components of variance*

STRUCTURED

- *Grouping or Nested—fungicides with related chemical or commercial structures*
- *Gradient or Regression—100, 150, 200ppm*
- *Factorial—each factor can have any of the previous four structures*

Other examples. Each of these treatment designs has some application in almost every discipline. Strains of biological material provide a common sort of treatment that often is regarded as unstructured. For example, you might be interested in cultivars of onions, breeds of cattle, species of field mice, strains of a pathogen, or races of wheat rust or of humans, etc. In sociological or behavioral studies, geographic locations may have no structure; in learning studies, different textual materials provide another example. Gradient treatment designs appear not only in regression studies, but also in many experiments that deal with substrate conditions such as temperature, light intensity, day length, water speed (for fish), fertilizer levels (for crops), level of feed consumption or dietary additives (for animals), amount of preconditioning (with humans), or length of instruction time (education), etc. In each, we seek ways to relate changes in the response of interest to changes in the condition.

Net effect. The treatment design focuses attention on certain comparisons among treatments — comparisons having particular interest to the researcher. It represents the "guts" of a research study; the other parts exist to support it. The treatment design dominates the important part of the statistical analysis. For the purposes of discussion,

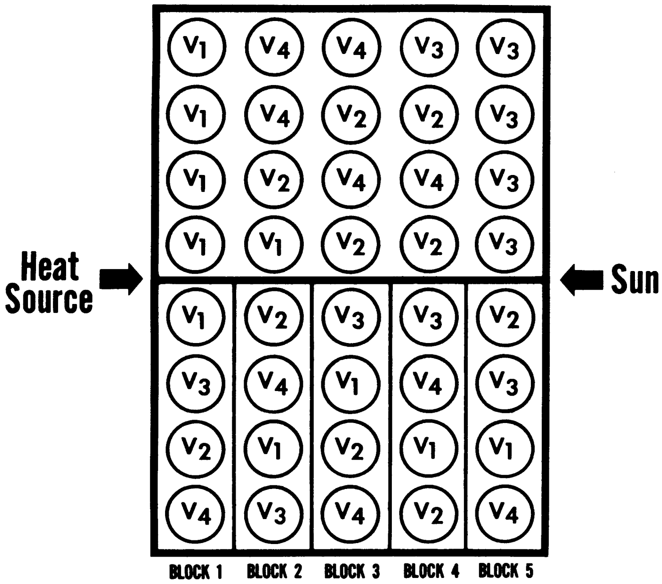

Fig. 1. Schematic of varieties appearing in a greenhouse.

suppose that we are in a situation where the analysis of variance is appropriate. Then, to examine how the treatment design would influence the analysis, we must first decide whether we have a treatment design involving a random set of levels of the treatment or a fixed set. Various variance components quantify the impact of the treatments on the experimental material for a treatment design involving a random set of levels.

If instead we have a fixed set of treatments, then various treatment means contain the information of interest. In other words, the treatment means start out as the fundamental parameters, but the treatment design suggests various ways in which they should be examined to maximize the relation between the analysis and the underlying research concern. Unstructured treatments cannot be examined relative to their structure because they have no structure. However, differences among treatment means can be assessed using multiple comparisons [LSD or Duncan's procedure; Cramer and Swanson (2)].

A structured set of treatments suggests various combinations of the treatment means (μ_i) that have a more specific relation to the research concern that the simple treatment means have. The treatment sum of squares in the analysis of variance supports an examination of the overall equality of underlying μ_i. Otherwise, it contains little information about how the μ_i relate to one another through the treatment design. The treatment design often suggests that an alternate set of parameters,

$$\eta_i = \sum_{j=1}^{t} c_{ij}\mu_j,$$

has an explicit meaning, so we should concentrate attention on them. These alternate parameters enable us to further break the treatment sum of squares into parts that relate specifically to the treatment design. The determination of the coefficients c_{ij} poses the major difficulty in this process, for once they are determined, associated calculations are described in most statistics textbooks under a heading such as contrasts or individual degrees of freedom. See for example, Steel and Torrie (14, section 11.8). In complex cases, the choice of the c_{ij} may require that you collaborate with a biometrician. See (9, 10, 17) for illustrations.

Experimental designs

An *experimental design* specifies the relation between the treatments and the experimental units. We can start by considering a random sample from each treatment population. Consequently, when treatments are applied to experimental units, in contrast to the case when the treatments represent a characteristic of the experimental units, such as a cultivar, the allocation of treatments to experimental units should be completely at random, unless there are good reasons to the contrary, as discussed in the sequel.

For example, if 5 pots each of 4 cultivars of tomatoes were randomly assigned to locations on a greenhouse bench, the layout in the top part of Fig. 1 might result. This represents a completely valid randomization, even though all of cultivar 1 falls on one end and all of cultivar 3 on the other end. However, a horticulturist might well inquire, "Suppose the heat source or the sun location has some effect. Couldn't this influence the response of cultivars 1 and 3 differently than of 2 and 4? That's why I don't like this randomization."

This does not argue against the randomization itself; rather, it says that *we should take subject matter knowledge into account at the randomization stage.* The influence of known experimental heterogeneity, such as thermal gradients, can be minimized by placing proper restrictions on the association of treatments with experimental units. In the present example, this would require that we separate the bench into 5 "blocks" of 4 pots each and restrict the randomization so that each cultivar appears once in each block, as shown in the bottom part of Fig. 1.

Recall again the chrysanthemum example concerning water sources. In that example, 72 groups of 5 pots each were placed in an experimental area, 24 groups on each of 3 benches. It has been established that plant responses vary somewhat across a greenhouse. If the location of experimental material influences the response, then, by restricting the location of various treatments relative to each other, we can reduce the bias or misleading effect of this extraneous variation. This mum example closely parallels the preceding tomato example.

Both of these examples provide an illustration of an experimental design that usually is called a *randomized complete block experimental design*. Other common experimental designs are called *completely random, latin square, split plot, balanced incomplete blocks, lattices,* and *partially balanced incomplete blocks*. Rather than describe these here, it suffices to note that these experimental designs cover a wide spectrum of research situations, and most of them are suited for use with any treatment design. See Cochran and Cox (3).

Net effect. The experimental design serves to minimize the effect of uninteresting but known sources of variability in experimental material. The experimental design has no effect on the analysis suggested by either the treatment design or the response design. It can affect how to estimate the treatment means before doing the analysis suggested by the treatment design, but it mainly affects methods for isolating valid estimates of residual variance (error).

Responses

Statistical considerations often begin by assuming the availability of a response; however, before launching into a consideration of kinds of responses or their relation to the experimental units, we should reflect on questions like "what constitutes a response?" or ". . . a *good* response?" Such questions defy precise answers because their answer depends heavily upon the area of investigation, but a few general remarks apply to all situations. The response should clearly characterize the attributes of interest in the population of experimental units. Specifically, if 2 experimental units differ in the attribute of interest, then the measured response should have different values for them.

EXPERIMENTAL DESIGNS

- *The EXPERIMENTAL DESIGN specifies how the treatments relate to the experimental units.*
- *Treatments should have a completely random association with experimental units in the absence of strong reasons to the contrary.*
- *Good reason: strata or blocks of homogeneous experimental units exist.*

Does the planned response really reflect what you want? In produc-

KINDS OF EXPERIMENTAL DESIGN

Popular Experimental Designs	*Restriction on Randomization*
completely random	*none*
randomized complete block	*once in each block*
latin square or latin rectangle	*two: rows and columns*
split plot	*two stages*

Covariance may be a good alternative to blocking if a meaningful covariate exists.

tion agriculture, the question of relevant responses often has a simple answer. If a grower gets paid on some unit basis, you can use the amount of this unit as the variable. Almost any other interest can lead to serious concern about what is the relevant variable. For example, how should you "measure" disease resistance of onions, quality of apples, or the impact of floral arrangements in nursing homes?

Fairly obvious responses emerge in many problems using physical responses, but problems occur even here, because most instruments detect some consequence of the concept and translate it into a response value, a process that works when certain assumptions apply. For example, Wilson (18, p. 45) discusses problems with a voltmeter actually producing a voltage reading. Recent work of the Growth Chamber Working Group (7, 15, 16) demonstrates how difficult it is to define and measure things as apparently obvious as temperature, CO_2 concentration, or (solar) radiation. Much more obvious illustrations occur when researchers must quantify concepts related to psychological or behavioral phenomena. Frankly, each discipline must face its problems of finding how to evaluate responses which reflect changes in its concepts.

How should you deal with the question of relevant responses? Begin by defining your interest in conceptual rather than operational terms. For example, suppose you are interested in the effect of concentrations of a chemical on a leaf spot disease. This is a conceptual statement. Contrast this to an operational statement: Score each plant (0–10) and average the score for 15 plants per plot; or use plot yield as the measure of the effect of the disease. At interpretation time you need to be very aware that your operationally defined (and analyzed) variable may fall short of your conceptual interest. By recognizing both conceptual and operational definitions you will be more aware of your variable's limitations. If this discussion seems incomplete, consider sections 4.1–4.6, 9.1 and 9.6 in Wilson (18).

Continuous and discrete responses stand as opposites of each other. Discrete responses can assume only isolated values usually associated with counts, while continuous responses can assume any value between 2 bounds.

RESPONSES

Principle: Responses (or measures or variables) should reflect on the object of interest — nothing more and nothing less.

e.g.: consider height of umble above ground
vs.
length of seed stalk.

Discrete responses. Two rather different situations produce discrete responses from counts. For example, in studying sex ratios we might classify insects as male or female; in plant breeding we might classify roses as red, pink, or white; in sociology we might classify people in the Southwest as English-speaking, Spanish-speaking, or bilingual. We could simultaneously classify responses by 2 or more characteristics, such as classifying people by language capability and socio-economic status, or dairy cows by udder-type and body-type, etc. After classification, discrete data result from counting the number of experimental units in each class or category. The binomial or various multinomial distributions underlie the statistical formulation and analysis of such categorical data. The other sort of discrete data likewise aise from counts, but without any categorization. Consider counting the deer on a plot of ground, the times a particular banded quail enters a bird traps during a summer, the adult males seeking jobs through a particular employment office, or the weeds in a particular farm plot. Numerous probability distributions exist for such responses. The Poisson and the negative binomial are 2 of the important ones. Sometimes you might mistakenly conclude that you have this latter type of discrete data when you really have a categorical situation. If your counts have a known upper limit, you probably have categorical data. For example, you may count the number of seeds that germinate in a plot, but if you know how many you planted, then you have the dichotomous response of germinate/not germinate, counts of which could behave like binomial random variables.

Continuous responses. Do continuous responses exist? Certainly. Consider your height: You began life as a zygote, with a height near (bounded below by) zero; as you grew to your present stature, you assumed every height between essentially zero and your current height.

Any numerical evaluation of a continuous response represents a discrete approximation in this sense: If we measure to the nearest 0.01, then a recorded response of 17.23 means that the actual response lay between 17.225 and 17.235, and similarly if we measure to the nearest 0.1, 0.001, etc. This discreteness of evaluation poses no difficulty provided we do not seek to detect differences between treatments of the same magnitude at this precision of measurement. Of course, with additional care, more precise instruments, more exacting conditions of evaluation, 0.01 can be reduced to 0.001, for example. Operationally, this observation serves as a criterion for testing the continuity of a response. Thus, if additional care, more precise instruments, etc., can produce increasing refinement in a response's evaluation, then you are considering a continuous response. This condition is only sufficient, not necessary, i.e., there exist continuous responses that we cannot evaluate. Most behavioral, sensory and many condition responses evade numerical evaluation because we lack an instrument for measuring the concept. For example, plants or animals differ perceptibly in their vigor. We can arrange several plants or animals of the same species in order by their vigor without ever getting a numerical value for "vigor"; or, in other situations, we could replace "vigor" by palatability for foods, by aggressiveness of humans, or endurance of fish, etc.

Relative and absolute evaluation. Thus, our evaluation of a continuous response may produce either relative or absolute responses. Ranks frequently get involved with relative responses, whereas absolute evaluation produces our usual "numbers". Constructed responses, unfortunately, may masquerade as absolute evaluations. Numbers, not measures, result from making arithmetic composites of various human perceptions or actions. For example, the landscape architect studying the asthetic appeal of a planting, the behavioral scientist studying a bird's territoriality, the child psychologist studying pre-schoolers' mathematical intuition, or the plant scientist studying disease resistance, all sense certain situations, assign numbers to various situations in some rational manner, and create a score by combin-

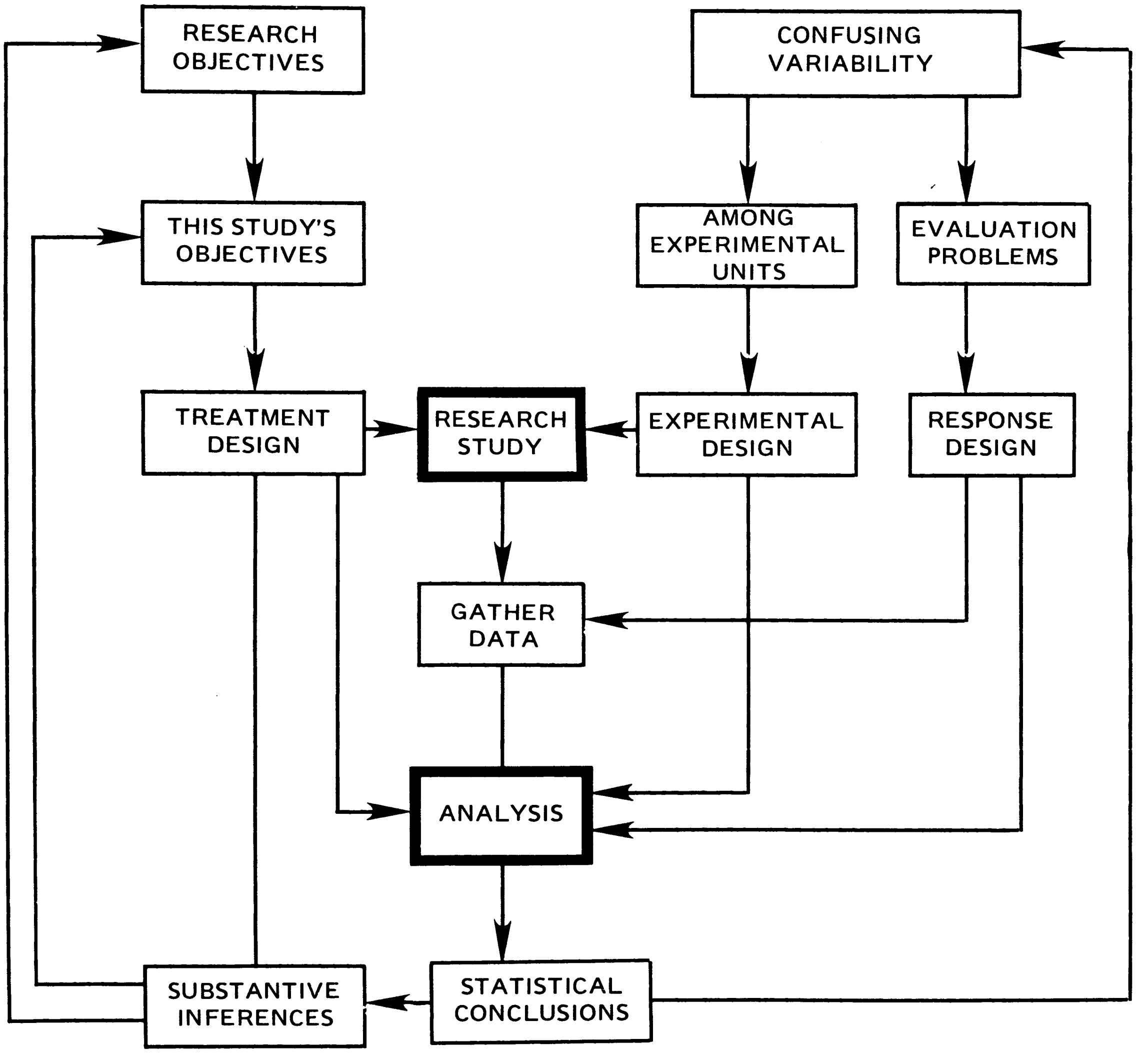

Fig. 2. Schematic diagram of a research study.

ing these numbers; again, probably in a rational fashion. But do these people have measurements? Certainly not! A kilogram is a kilogram whether it represents the unit difference between 1 and 2 kg. or between 1000 and 1001 kg.; the same thing cannot be said about a created score. For example, a plant with an infection score of 4 suffers the disease more severely than one with a score of 3, but both remain fairly healthy, while the same unit difference between scores of 9 and 10 could mean the difference in life or death of the plant. As a valid score increases, the response we seek to characterize increases, and conversely. (We could use this property to examine the validity of a scoring scheme.) In this way a score forms the basis for comparing experimental units relative to the responses by producing a relative, (not absolute) evaluation of the response.

An even more confusing situation occurs when we use an absolutely evaluated response to characterize some other response. For example, to study how several rearing situations (treatments) could affect birds' propensity to disperse, we could raise birds under the several situations, release them from some common point and measure their distance from the release point after some reasonable time. Distance has a scale with well-defined units, but does 1 distance unit correspond to 1 unit of propensity to disperse? Probably not. Consider 2 birds that went 0.1 and 1.1 miles, as opposed to 2 birds that went 24 and 25 miles respectively. I am inclined to say that the first pair differed greatly in their propensity to disperse, while the second differed little, even though, 1.1–0.1 = 1.0 = 26–25.

You must face the problem of whether your data — numbers — represent scores or the direct evaluation of a response of interest on a well-defined scale. This distinction is important, because the statistical formulation of your experimental questions and the associated analyses rest heavily upon your decision. Comparatively evaluated responses usually should be analyzed by rank techniques, while absolutely evaluated responses often yield to the analysis of variance or regression techniques.

Continuous responses, determined with the assistance of some device (such as a length or a weight), enable us to examine further issues, namely, precision and accuracy. Precision deals with the closeness together of several evaluations of the same response, for example, several weighings of the contents of a package of sugar. Special procedures must be invoked if you want to detect differences between treatments that are of the same magnitude as your precision. Accuracy, on the other hand, deals with what we loosely call bias. Any miscalibrated instrument provides a superficial, but nontrivial, example. You should check both precision and accuracy on usual as well as new responses.

TYPES OF RESPONSES

- *Discrete*
- *Continuous*

  Ordinal = rank — Relative Evaluation

  Measure — Absolute Evaluation

Response designs

The *response design* describes the relation of individual responses to their experimental units. In the mum example, each experimental unit contained 5 pots from which plant heights were obtained. These individual plants provide an example of what we will call evaluation units.

Evaluation units. We seek inferences from populations of experimental units, but we may have difficulty evaluating responses on the experimental units. Laboratory evaluation of chemical responses provides especially obvious examples. We cannot practically evaluate the calcium content of a bale of hay, the caloric equivalent of a sheep's body, or the size of each egg in a female trout ready to lay eggs. Similarly, to study the grammatical style of a particular author we may make a detailed examination of a sample of pages from his writing. A dairy scientist trying to evaluate the breeding potential of bulls for transmitting milk production cannot even get a related response from the experimental units; he must study the bulls' daughters' milk production.

These examples suggest an obvious way to define an *evaluation unit,* namely that unit of experimental material that yields a response. At first, this definition of an evaluation unit may seem very similar to other authors' definitions of sampling unit as a randomly selected part of an experimental unit. But an evaluation unit exhibits greater generality in 2 ways. It does not have to represent a random sample, because sometimes we will allow the unit to have a structure. Further, the evaluation unit may be physically separate from the experimental unit: witness the bulls' daughters.

Repeated evaluations. When some difficulty attaches to the evaluation of the response on an experimental unit, such as measuring the length of a wiggling fish, we can make several *independent* evaluations of the response to increase the measurement precision. This has a very minor effect on the analysis. On the other hand, *nonindependent* repeated evaluation can substantially alter the analysis. Numerous examples of this occur when experimental units are subjected to treatments for some time, like plants given a nutrient, animals on diets, or students in a learning situation, because the response can be evaluated several times.

Net effect. Normally, we seek inferences to populations of experimental units. Thus, most of our interpretation rests upon composites of responses from the evaluation units — one composite for each experimental unit. We can draw inferences from the response design, but these inferences tell us about our evaluation process, nothing about the experimental unit or treatments. They have great value in planning a good response design in future studies, but otherwise, relatively few subject matter inferences come from the analysis of the response design.

Relation of the Parts

We have considered 3 major parts of a research study: 1) the response design, 2) the treatment design, and 3) the experimental design. Now consider how these relate to each other. Fig. 2 relates these parts schematically.

Impact on the study. Your research objectives translate into specific objectives for a study; these dominate your choice of treatments. Once the treatments have been selected, they exert no influence on the research study other than their presence. Confusing variability manifests itself in variation among experimental units and separately in evaluation of responses from them. The experimental design is chosen to minimize the impact of variation among experimental units; it has rather substantial impact on the setup and conduct of the study. The response design first impinges on the study when the data are gathered, although it should be thoroughly planned before the study is started.

Source of Variation	Degrees of Freedom
Total	*tev*
Mean	*1*
Among Evaluation Units	*tev-1*
Among Experimental Units	*te-1*
Among Treatments	t-1
⋮ } Directed by the Treatment Design	⋮
Among Exp. Units Within Treatments	*t(e-1)*
⋮ } Directed by the Experimental Design	⋮ } h
Residual 1	t(e-1)-h
Among Eval. Units Within Exp. Units	*te(v-1)*
⋮ } Directed by the Response Design	⋮ } s
Residual 2	te(v-1)-s

Fig. 3. A general analysis of variance table demonstrating the effect of the treatment, experimental and response design.

Source of Variation	Degrees of Freedom
Total	*360*
Mean	*1*
Among Individual Pot Heights	*359*
Among Groups of Plants	*71*
Among Water Sources	23
Regression	k
Lack of Fit	23-k
Among Groups of Pots Within Water Sources	*48*
Benches (= Blocks)	2
Residual	46
Among Pot Heights Within Groups	288

Fig. 4. Specific analysis of variance table for the chrysanthemum illustration.

Impact on the analysis. Each design component influences the analysis in a fairly distinctive way. For the present, consider illustrating this with an analysis of variance. Fig. 3 and 4 present a general analysis of variance and illustrate it with the mum example. Only sources of variation and degrees of freedom appear; the other columns (sums of squares, mean squares, expected mean squares and F-statistics) depend to varying extents on the specific nature of the 3 design components. For symbolism, suppose the research study has *t* treatments, *e* experimental units per treatment, and *v* evaluation units per experimental unit. First, total variation is partitioned into the mean (for over-all size of the response) and variation among (or between) evaluation units; this is decomposed into variation among experimental units and the remainder, which is variation among evaluation units within experimental units. (Note that varying indentations indicate different levels of subdivision of sums of squares.) Variation among experimental units is partitioned into its 2 contributory components: among treatments and among experimental units within treatments. At this point the 3 design components take over so that the completion of the partitioning depends on them.

The italicized lines in these analysis of variance tables usually will not appear in a completed table because they represent only intermediate steps. The total line may appear without the mean line, indicating a "corrected" total, i.e., a line arrived at by taking the difference between the sums of squares and degrees of freedom of the first 2 lines indicated here. We will include both lines rather than their difference, because this explicitly accounts for all of the degrees of freedom.

The above discussion of analysis was oriented toward the analysis of variance, a technique designed for measurement (not relative) data. Earlier, we noted the existence of discrete and ordinal types of data. Analysis-of-variance-like procedures are being developed for such situations. See Grizzle, Starmer, and Koch (6) for discussion of the conceptual background of these procedures for discrete data; they are illustrated in Koch (9). Powerful statistical techniques also exist for the analysis of rank data. See Conover and Iman (4) for an entry point into this literature.

Comments on terminology

The users of statistics encounter a frustrating problem: statisticians seem inconsistent in the definitions they attach to certain words and in their use of symbols. This problem occurs for 2 rather good reasons.

First, statistics is a rather young discipline; as such, it has not gone through the standardization process that has been experienced by chemistry, physics, and many of the more-established sciences. Further, what an author tries to communicate to a particular audience influences his choice of notation and terminology. In many instances, several reasonable terms may describe an idea, but an author ordinarily chooses the term his audience is most likely to understand. Thus, we will find that an author writing for agriculture may use one set of names for something, whereas an author writing for education may use another, while engineers use still another set of names.

RESPONSE DESIGNS

Definition: An evaluation unit is the unit of research material on which a response is evaluated.

e.g. Experiment Unit = plot of onions
Evaluation Unit = seed stalk (height)

- *The response design specifies the relation between the experimental units the evaluation units, and the responses.*

Originally, much of what now is regarded as standard statistical procedures came about to meet problems in agriculture. Many of the names that were then associated with basic ideas in statistics seemed unreasonable, and in a strictly agricultural situation, they remain reasonable. However, as statistics has found application in an expanding set of areas, these names become less and less appropriate. In writing this anatomy, I sought to develop some general ideas about the meaning of some basic terms. I chose names that should communicate an idea; in doing this I coined 2 new phrases and expanded meanings of some common phrases to the general situation.

For example, the phrase "experimental design" appears to have 2 meanings in the statistical literature. In one sense it deals with all aspects of the analysis of data through the statistical technique called the analysis of variance. In these settings, exemplified in various books (3, 5, 8, 11, 13), the word "experimental design" usually remains undefined. If it is defined, the definition is cursory. In other writings, where authors define an experimental design, it has a specific meaning, related fairly closely to mine — that is, the relation between the treatments and the experimental units. For example, Ostle (12) and Steele and Torrie (14) illustrate this view of experimental design.

An experimental unit usually is defined as the amount of experimental material to which an individual treatment is applied. This definition fits production agriculture, but does not apply effectively to many areas of research. Frequently, treatments do not actually get applied to the experimental unit, but come with it; for example, breed of a cow or the age of a child. The experimental unit is selected from a population of objects with that characteristic. The definition of experimental unit set forth earlier represents a broadening of the usual definition to cover such situations.

The treatment design, which specifies how the treatments relate to each other for the purpose of drawing context inferencess, frequently gets recognized implicitly by various authors, but it does not become a central focus. In many statistical writings, chapters and sets of chapters are organized around distinction between the treatment design and experimental design. Only Federer (5) appears to introduce the idea of a treatment design explicity, but he merely mentions it. From conversations with him, I know now it has become central in his thinking.

The response design usually gets relegated to a secondary position in much of agriculture and biology. Thus it does not appear in writings of statistical topics related to these areas. On the other hand, it does get recognized, though not usually named explicitly, in some of the behavioral sciences. For example, the book by Myers (11) has 3 chapters structured around this idea.

In conclusion, remember that your interest lies in research, not in statistics. Nevertheless, statistics can help you conduct and analyze experiments so that you will obtain the most information possible for fixed resources. To accomplish this maximization you need to: carefully state your objectives; choose responses (measures and treatments) which will reflect on the objectives; set up the experimental material to block out known sources of heterogeneity; tune your statistical analysis to the treatment, experimental, and response designs used; and interpret the statistical results in relation to your objectives.

Literature Cited

1. Box, G. E. P., N. R. Draper, and J. S. Hunter. Empirical model-building with response surfaces. Wiley, New York. (In press)
2. Carmer, S. G. and M. R. Swanson. 1973. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. J. Amer. Stat. Assoc. 68:66–74.
3. Cochran, W. G. and G. M. Cox. 1957. Experimental designs, 2nd ed. Wiley, New York.
4. Conover, W. J. and R. L. Iman. 1981. Rank transformations as a bridge between parametric and non-parametric statistics. Amer. Statistician 35:124–133.
5. Federer, W. T. 1955. Experimental designs: theory and application. Macmillan, New York.
6. Grizzle, J. E., C. F. Starmer, and G. G. Koch. 1969. Analysis of categorical data by linear models. Biometrics 25:489–504.
7. Hammer, P. A., T. W. Tibbitts, R. W. Langhans, and J. C. McFarlane. 1978. Baseline growth studies of 'Grand Rapids' lettuce in controlled environments. J. Amer. Soc. Hort. Sci. 103:649–655.
8. Kempthorne, O. 1952. Design and analysis of experiments. Wiley, New York.
9. Koch, G. G., J. E. Grizzle, K. Semenya, and P. K. Sen. 1978. Statistical methods for evaluation of mastitis treatment data. J. Dairy Sci. 61:829–847.
10. Little, T. M. 1981. Interpretation and presentation of results. HortScience 16:637–640.
11. Myers, J. L. 1966. Fundamentals of experimental design. Allyn & Bacon, Boston.
12. Ostle, B. 1963. Statistics and research, 2nd ed. Iowa State Univ. Press, Ames.
13. Scheffe, H. 1959. The analysis of variance. Wiley, New York.
14. Steel, R. G. D. and J. H. Torrie. 1960. Principles and procedures of statistics. McGraw-Hill, New York.
15. Tibbitts, T. W. 1978. Standardization of controlled environment research. HortScience 13:451.
16. Tibbitts, T. W., J. C. McFarlane, D. T. Krizek, W. L. Berry, P. A. Hammer, R. W. Langhans, R. A. Larson, and D. P. Ormrod. 1976. Radiation environment of growth chambers. J. Amer. Soc. Hort. Sci. 101:164–170.
17. Urquhart, N. S. and D. L. Weeks. 1978. Linear models in messy data: some problems and alternatives. Biometrics 34:696–705.
18. Wilson, E. B. 1962. An introduction of scientific research. McGraw-Hill, New York.