The Entity-Property-Relationship Approach to Statistics: An Introduction for Students

Donald B. Macnaughton*

This paper describes the "entity-property-relationship" approach to science and statistics at a level suitable for students in an introductory statistics course.

1. INTRODUCTION

This paper discusses the vital role of the field of statistics in scientific research. Discussion is in terms of the "entity-property-relationship" approach (Macnaughton 1996a). Many students enjoy learning about the approach because it is complete, yet easy to understand.

2. ENTITIES

If you stop and observe your train of thought at this moment, you will probably agree that you think about various "things". For example, during the course of a minute or so you might think about, among other things, a friend, an appointment, today's weather, and an idea. Each of these things is an example of an *entity*.

Many different types of entities exist. For example, the following diverse types of "things" are simply different types of entities:

- physical objects (e.g., mountains, automobiles, protons)
- processes (e.g., chemical reactions, stage plays)
- organisms (e.g., people, trees)
- events
- thoughts (e.g., theories, concepts, ideas, emotions)
- societal entities (e.g., educational, political, and commercial institutions)
- symbols
- forces (e.g., force needed to lift a physical object, magnetic forces)
- waves (e.g., water waves, sound waves, electromagnetic waves)
- mathematical entities (e.g., numbers, sets, elements of sets, functions, vectors).

Entities are fundamental units of human reality because everything in human reality is an entity.

Entities play a central role in language: a fundamental part of speech—the noun—is used solely to represent entities.

People usually view entities as existing in two different places: in the external world and in their minds. We use the entities in our minds mainly to stand for entities in the external world, much as we use a map to stand for its territory.

Most people begin to use the concept of *entity* when they are very young. Most of us use the concept automatically as a way of organizing the multitude of stimuli that enter our minds minute by minute when we are awake.

Because people use the concept of *entity* almost entirely at a subconscious level, some people have difficulty grasping the fundamental role that the concept plays in their thought.

The concept of *entity* is further concealed because it is not often directly discussed in science or statistics. Direct discussion is usually omitted because, in dealing with specific issues, it is not often necessary to drill down all the way to the foundational concept and discuss "things" at such a basic level. Instead, discussions usually concern one or more particular *types* of entities, which are best referred to by their type names. For example, medical scientists often study a type of entity called *human beings.* However, as we shall see, the concept of *entity* serves as the foundation for other important scientific and statistical concepts, and therefore it is helpful to make the concept explicit.

At those times when entities *are* discussed in general terms in science or statistics, they may also be called *cases, items, individuals, instances, specimens, subjects, (experimental) units, survey units,* or *members of the population.*

3. PROPERTIES OF ENTITIES

Every entity has associated with it a set of attributes or *properties.* Table 1 lists some entity types and some of the properties associated with entities of each type.

Properties are an important aspect of entities because we can only know or experience an entity by (in a broad sense) knowing or experiencing its properties. For example, if we experience a new drinking cup, we experience it through its properties of weight, color, shape, and volume.

The broadness of the concept of a property is reflected in the number of different names by which the concept is known. That is, properties are sometimes called *aspects, attributes, characteristics, characters, factors, features, qualities, quantities, scalars,* and *traits.*

In science, properties of entities are often called *variables.* That is, when scientists or statisticians refer to a variable, they are usually referring (either specifically or generally) to some property of some type of entity.

Kendall, Stuart, and Ord (1987, sec. 1.1-1.3) further discuss the central role that the concept of a property of an entity plays in the field of statistics.

4. VALUES OF PROPERTIES OF ENTITIES

For any particular entity, each of its properties has a *value.* We usually report the value of a property with words, with symbols, or with numbers. For example, table 2 lists some of the properties and the associated values for the entity known as the United Nations Building in

---

*Donald B. Macnaughton is president of MatStat Research Consulting Inc. 246 Cortleigh Blvd., Toronto Ontario, Canada M5N 1P7. E-mail: *donmac@matstat.com* Portions of this material were presented at the Joint Statistical Meetings in San Francisco, August 11, 1993 and at the Joint Statistical Meetings in Orlando, August 15, 1995. The author thanks Professor John Flowers of the University of Toronto and Professors Donald F. Burrill and Alexander Even of The Ontario Institute for Studies in Education for testing the approach in their statistics courses. Their comments and their students' comments helped substantially to clarify the ideas. The author also acknowledges insightful comments from W. Edwards Deming, Olaf E. Kraulis, Alexander M. Macnaughton, Thomas L. Moore, and Milo A. Schield.

Table 1. Some Entity Types with Examples of Some of Their Properties

Entity Type	*Properties of Entities of this Type*
physical objects	weight chemical composition age
human beings	height blood type intelligence quotient political affiliation whether presently alive
forces	magnitude direction locus of application
national economies	gross national product cost of living rate of inflation
events	probability of occurrence whether occurred duration
populations	size proportion of the population having a specified level of a property
works of art	beauty

New York City.

In everyday language, people use adjectives and adverbs to report the values of properties. For example, a person might use the adjective *tall* to report (the value of) the height (property) of a building or the adverb *quickly* to report (the value of) the speed (property) of (the process of) someone running in a race.

Adjectives and adverbs are useful for reporting the values of properties because they are compact—within a single word we can both identify the property of interest and indicate a particular value of it. However, adjectives and adverbs are usually imprecise. (How tall is *tall?*) If we need higher precision in the report of the value of a property, we can use numbers because numbers can represent any degree of precision we wish.

If we need to *know* the value of a property of an entity, we can apply an appropriate measuring instrument *to* the entity. If the instrument is measuring properly, it will return a measurement to us that is an estimate of the value of the property in the entity at the time of the measurement. For example, if we need to know the (value of the) height (property) of a person, we can apply a height-measuring instrument (e.g., a tape measure) to the person, and the instrument will give us a number that is an estimate of the person's height.

Table 2. Properties of the United Nations Building and Their Associated Values

Property	*Value of the Property*
height	tall (i.e., the word *tall*)
height in meters	165.8
primary building materials	concrete, glass, steel

Most people make and consider many references to values of properties of entities every day. We usually handle these references automatically, without being aware that we are using the general concept of the value of a property of an entity. Thus the concept of the value of a property of an entity is a fundamental concept of human thought.

To most people, the concept of the value of a property of an entity is intuitive. However, it is helpful make the concept explicit in an introductory statistics course because explicit discussion helps to clarify the key follow-on concept of a relationship between properties.

5. EXERCISES

1. Name five important types of entities in your favorite area of science, commerce, politics, recreation, law, technology, religion, art, or other field of interest. Name three properties of each type of entity. Describe briefly how each property can be measured.
2. In physics, an important type of entity is the entity *physical object.* Most of us believe that physical objects visible to the naked eye are made up of small sub-entities called molecules (which are also physical objects). These, in turn, are made up of sub-sub-entities and so on. Name some sub-entities of the five types of entities you named in exercise 1. Name some sub-sub-entities and so on as far as you can go. Name some properties of each type of sub-entity.
3. Obtain a rectangular piece of paper and a ruler. One property of the piece of paper is its length along each of its four edges. Choose one of the edges and measure the length with the ruler as accurately as you can. Measure the length along the same edge two more times to give you, altogether, three estimates of the length of the paper. How accurate are your estimates?
4. Do all measured values contain errors (large or small) or do some measuring instruments make "perfect" measurements? What would it mean to make a perfect measurement of the value of a property?

5. Some properties of entities can be measured in more than one way. For example, the temperature of an environment can be measured with a mercury thermometer, an alcohol thermometer, a bimetal thermometer, a thermocouple thermometer, an optical pyrometer, and with the volume of a fixed mass of gas maintained at a constant pressure. Name some other properties of entities that can be measured in more than one way and list each way you know of measuring these properties.
6. Name some highly accurate measures of properties of entities. Name some relatively inaccurate measures that are used to measure properties of entities.
7. List some ways that errors in measurement of the values of properties of entities can occur.

6. A GOAL OF SCIENCE: TO PREDICT AND CONTROL THE VALUES OF PROPERTIES

An important goal of science is to discover how to accurately *predict* and *control* the values of properties of entities. In other words, the goal is to predict and control the values of *variables in* entities. For example, a goal of medical science is to discover how to accurately predict and control the state of the human body, where the state is reflected by various medical properties or variables, such as blood pressure, white blood count, and other indicators of health or disease.

Society supports science in its search for the ability to predict and control the values of properties because instances of such ability often provide substantial social or commercial benefits. For example, if medical researchers can discover how to better predict and control a person's propensity to heart attacks, the discovery will provide the social benefit of saving lives.

We shall discuss the role of statistics in scientific research in terms of the general scientific goal of accurate prediction and control of the values of properties of entities.

7. RELATIONSHIPS BETWEEN PROPERTIES (RELATIONSHIPS BETWEEN VARIABLES) AS A KEY TO SCIENTIFIC PREDICTION

7.1 Science as the Study of Relationships Between Properties

Given the goal of predicting and controlling the values of properties of entities, an obvious question is *How* can we predict and control the values of properties? The answer is by studying *relationships between* properties—that is, by studying relationships between variables.

For example, medical scientists have discovered that a relationship exists between the amount of fat ingested by a person and the probability that a person will have a heart attack. Roughly speaking, the more fat a person ingests, the higher the probability that the person will have a heart attack. This relationship helps doctors and patients to predict and control heart attacks.

The concept of a relationship existing between properties of entities pervades all branches of science. One way to see this pervasiveness is to examine the so-called "laws" of science. It turns out that we can view most of these laws as statements of relationships (or occasionally non-relationships) that exist between properties of entities.

For example, an important law of physical chemistry is the ideal gas law,

$$pV = nRT.$$

This law relates pressure (p), volume (V), amount (n), and temperature (T) of a quantity of an ideal gas. (R is the constant of proportionality.) This law is a statement of a relationship between properties. It tells us that, if we know the values of any three of the properties for a quantity of an ideal gas, we can correctly predict the value of the fourth.

Similarly, Einstein's equation,

$$E = mc^2,$$

is a statement of a relationship between two properties of a piece of matter: energy of mass (E) and mass (m). (The c^2 can be viewed as the constant of proportionality, which Einstein has shown to be equal to the square of the speed of light.) If we know the mass of a piece of matter, we can use Einstein's equation to predict its energy of mass, and vice versa.

Examination of different branches of science suggests that many of the important statements within each branch of science (whether physical, biological, or social) can be viewed as statements of relationships between properties of entities (relationships between variables).

The concept of a relationship existing between properties extends beyond science to many areas of life, ranging from the simplest of aphorisms (e.g., haste makes waste) to the most complex of abstract concepts.

Mathematical equations are fundamental statements of mathematics. When scientists or mathematicians use a mathematical equation to reflect a state of affairs in the external world (as opposed to using an equation to make an abstract mathematical statement), the equation can almost always be viewed as a statement of a relationship between properties of one or more entities.

We can describe a relationship between properties or variables in general terms as follows:

A *relationship* exists in entities between a property y and one or more other properties $x_1,\ldots,x_p$ if any of the following (equivalent) conditions are satisfied:

- the measured value of y in the entities "depends" on the measured values of the xs in the entities or
- the measured value of y in the entities "varies" wholly or partially "in step" with the measured values of the xs in the entities or
- y is some *function* of the xs in the entities—that is

$$y = f(x_1, x_2, \ldots, x_p) + \varepsilon$$

where $f(\bullet)$ is a mathematical function and ε represents random variation.

In summary, we can study relationships between properties in any of the multitude of different types of entities in the world around us. And armed with knowledge of relationships between properties obtained from such study, we can, with some accuracy, predict and sometimes control the values of properties. This conceptually simple but surprisingly powerful procedure is central to all branches of science.

7.2 Exercises

1. The numbered list below names fifteen pairs of properties of various types of entities. For each pair of properties in the list, answer the following questions:
   (a) What is the name of the type of entities under consideration?
   (b) Do you think a relationship exists between the two properties in the entities?
   (c) If you think a relationship exists, do you think the relationship is strong, of medium strength, or weak?
   (d) If you know an equation that describes the relationship, state the equation.

Fifteen Pairs Of Properties

i. the weight of a car and the gas mileage of a car
ii. the education of a person and the income of a person
iii. the net amount of force applied to a physical object and the rate of acceleration of an object
iv. the amount of a particular substance used in a chemical reaction and the rate of an instance of the reaction
v. the thickness of the insulation in the walls of a building and the cost to heat a building
vi. the number of cigarettes smoked by a person in the last two years and the probability that a person will be diagnosed as having lung cancer in the next four years
vii. the amount of a headache drug taken by a person with a headache and the amount that a headache changes in severity
viii. the concentration of alcohol in the bloodstream of a person driving a car and the probability that a person driving a car will be involved in an automobile accident
ix. the style of child-rearing used with a child and the probability that a child will engage in criminal activity in later life
x. the style of management of a business and the volume of profit of a business
xi. the amount of money spent advertising a product and the volume of sales of a product
xii. the failure rate of a product and the volume of sales of a product
xiii. the number of hours of homework completed by a student during a course and the grade obtained by a student in a course
xiv. the identity of the jockey riding a racehorse in a race and the probability that a horse will win a race
xv. a person's astrological sign and traits of a person's personality.

2. For the entity types you listed in exercise 1 in section 5, list as many of the known relationships between properties of the entities as you can. For each relationship, state whether you think it is strong, of medium strength, or weak.
3. Name some areas of science in which most of the known relationships between properties are strong. Name some areas in which most of the known relationships are weak.
4. Suppose you are a horticulturist and you have a good instrument for measuring soil moisture content and a good instrument for measuring the healthiness of plants. Discuss how you would determine the relationship between soil moisture content and the healthiness of some specific species of plant.
5. An advertisement for an exercise machine is headed "No Pain, No Gain." In which areas of life do you think there is a relationship between pain and gain? In which areas do you think there is none?

7.3 Properties as Variables

The preceding material emphasizes the concept of *property* while alluding to the closely related concept of *variable.* In the rest of this paper (except at the highest level) we shall reverse the emphasis and discuss properties and relationships between properties in terms of variables and relationships between variables. The use of dual terminology, although cumbersome, is necessary because properties and variables both play important roles. Specifically, the concept of *property* is intuitive to most people, and thus use of the concept helps people to understand the ideas. On the other hand, the more technical concept of *variable* is entrenched at a fundamental level throughout science and statistics and is therefore mandatory for discussion.

Some beginning students understand the concept of *property,* but have trouble with the concept of *variable.* These students may find it helpful to remember that the variables discussed in science and statistics are simply "instances" of properties of entities, and that the *value* of any variable represents the value of the property in the associated entity (usually at a particular time).

Barnett (1988) gives a general discussion of the con-

cept of a relationship between variables. Macnaughton (1996b) discusses the distinction between properties and variables.

7.4 Population

After considering the key concept of a relationship between variables, it is next useful to consider exactly "where" scientists study such relationships.

Formally, scientists begin the study of a relationship between variables by defining a group of similar entities within which (or within whom) they wish to study the relationship. Scientists call this group of entities the *population* of interest. For maximum generality, scientists usually define a population to be all the entities in the world (or universe) of some particular type.

For example, medical scientists studying some particular disease are often interested in studying relationships between variables in the population of all the people in the world (ever) who have that disease. Similarly, physical chemists studying the ideal gas law are interested in studying relationships between certain variables in the population of all the instances (ever) in which a gas is held in a container.

Scientists are usually interested in studying relationships between variables *as these relationships exist in all the entities in a population.* Scientists seek relationships in all the entities in a population (as opposed to seeking relationships in individual entities, or in small groups of entities) because reliable information about a relationship between variables in all the entities in a population allows us (i.e., society) to make predictions or exercise control (with some accuracy) for any entity in the population.

7.5 Sample

Because populations are usually not fully accessible, scientists usually cannot examine every entity in a population for information about a sought-after relationship. Instead, scientists examine a subset of the population, which they call the *sample.*

For example, in a medical research project aimed at studying a new treatment for cancer the sample might be a group of patients in a particular hospital who have that disease. Similarly, in a research project aimed at studying the ideal gas law the sample might be several instances of amounts of gases being held in containers in a laboratory under various conditions of temperature, pressure, and volume.

In scientific research we examine the entities in a sample for evidence of a relationship between the relevant variables. (Details of how to do the examination are given below.) If reasonable evidence of a relationship is found (and assuming certain pitfalls have been avoided), we can then confidently generalize the relationship to all the entities in the population. Then we (as society) can (when socially appropriate) use our knowledge of the relationship to predict or perhaps control the values of certain of the studied variables for any entity in the population.

7.6 Response Variables and Predictor Variables

When studying a relationship between variables, a scientist will often divide the variables that are measured in a research project into two groups: response variable(s) and predictor variable(s).

Response variables are the variables that the scientist wishes to discover how to predict (or perhaps even control). *Predictor* variables are the variables that the scientist will measure (or perhaps even control) in the entities in the sample in an attempt to discover a relationship that will then enable us to predict (control) the values of the response variables in other entities in the population.

For example, when medical scientists study the relationship between "propensity to heart attacks" and "fat intake" in humans, they usually view "propensity to heart attacks" as the response variable, and they usually view "fat intake" as the predictor variable. They view the two variables this way because they usually wish to learn how to predict (control) the propensity to heart attacks by measuring (controlling) fat intake, not the other way around.

Scientists often find it efficient to concentrate their attention in a research project on learning how to predict or control the values of a single variable in the population of entities they are studying. Therefore, without loss of relevant generality, we restrict discussion to research projects (or separate units of analysis in research projects) that have only a single response variable.

On the other hand, our discussion will cover research projects that may have any number (including zero) of predictor variables.

The broadness of the concepts of response and predictor variables is reflected in the number of different names by which these concepts are known. That is, response variables are sometimes called consequent variables, criterion variables, dependent variables, effect variables, observed variables, outcome variables, output variables, predicted variables, and y -variables. Similarly, predictor variables are sometimes called (with various shades of meaning) active variables, antecedent variables, carrier variables, carriers, cause variables, classification variables, concomitant variables, control variables, covariables, covariates, explanatory variables, factors, grouping variables, independent variables, input variables, manipulated variables, predicated variables, predictors, regressor variables, regressors, stimulus variables, stratification variables, supplementary observed variables, treatment variables, and x -variables.

7.7 Exercise

1. For each pair of fifteen pairs of properties listed in exercise 1 in subsection 7.2, suppose you are planning a research project to study the possible relationship between the two properties. For each pair of properties in the list, answer the following questions:
   (a) In a research project to study the relationship between the two properties, which of the two properties would normally be the response variable?
   (b) What is the population?
   (c) Using your intuition, what do you think would be a reasonable number of entities to sample from the population to allow study of the relationship between the two properties and to allow generalization of any findings to all the entities in the population? (Of course, you will wish to avoid using "too many" entities in the sample because that will unnecessarily increase the cost of the research project.)
   (d) Using your intuition, how would you go about choosing the entities in the sample from all the entities in the population in order to ensure good generalizability (while, again, trying to keep down the cost of the research project)?

7.8 Scientific Research Projects

The list in exercise 1 in subsection 7.2 names fifteen possible relationships between properties. We can study each of these possible relationships (and relationships between any other pairs or larger sets of properties) in the context of a scientific research project. Each research project can be viewed in terms of

1. a population of entities
2. a sample of entities that is selected from the population
3. one or more predictor variables that are measured or controlled in each entity in the sample (or possibly are measured or controlled in each entity's environment)
4. a response variable that is also measured in each entity in the sample
5. (most importantly) a relationship in the entities in the population between the response variable and the predictor variable(s) that is sought or studied in the research project.

By considering a broad set of examples, one can see that most scientific (i.e., empirical) research projects can be usefully viewed as being studies of relationships between properties of entities (relationships between variables). That is, we can cast most scientific research projects in terms of the five points given in the preceding paragraph. This approach provides a simple yet comprehensive model of most scientific research.

Learning to view most scientific research projects as studying relationships between properties lays a broad and deep foundation for discussion of the role of statistics in scientific research, which we discuss below in sections 8 through 13.

7.9 Exercises

1. Consider a scientific research project with which you are familiar.
   (a) What is the population of entities that was studied?
   (b) How big was the sample, and how was the sample selected from the population?
   (c) What was the (main) response variable? What was (were) the predictor variable(s)?
   (d) Draw a graph that summarizes the (main) relationship between variables sought or studied in the research project.

NOTE: The following three exercises and some later exercises require the use of a statistical software package. Your instructor will introduce you to the computer system and the statistical package you will use for these exercises.

2. An education researcher wished to know if it is possible to accurately predict a student's grade in a course from the number of hours of homework that the student does in the course. Therefore, the researcher measured the number of hours of homework students did in a particular course and the grades of the students in the course for a group of 30 students. This yielded table 3. The researcher obtained the students for participation in the research project by simply using all the students in a particular university course that he was teaching.
   (a) In this research project what is the population of entities under consideration? Answer: Formally, the population in a research project is all the entities that theoretically have a chance of being selected for participation in the research project. Informally, the population is sometimes viewed as being much broader than the formal population, both in physical extent and in time. Generalizations to broader populations are sometimes reasonable. However, to avoid later possible embarrassment, generalizations to broader populations should be done with care.
   (b) What is the response variable in this research project?
   (c) Scan the data in the table. From scanning the data, do you think that there is a relationship between the two variables? If so, describe the relationship.
   (d) Enter the data in the table (all three columns) into the computer and have the statistical package print the data on the printer. Proofread the printout against table 3 to verify that you have entered the data correctly. When defining the data for the statistical package specify "long names" for the predictor and response variables. Generally, variable names should contain the name of the property that was measured in the entities and (if applicable) the

Table 3

Student ID Number	*Hours of Homework Done*	*Course Grade Attained*
1	46	66
2	1	60
3	18	46
4	15	59
5	101	81
6	116	80
7	67	68
8	71	66
9	54	67
10	79	67
11	14	57
12	41	70
13	83	68
14	59	67
15	58	65
16	34	69
17	48	61
18	129	82
19	98	80
20	10	64
21	56	63
22	132	88
23	22	51
24	66	77
25	83	75
26	58	61
27	1	57
28	17	63
29	110	79
30	43	70

units in which the property was measured. For example, for the variable reflecting the amount of homework done by students in an education research project, you might use a short variable name of "HMWORKHR" and a long name of "Homework Done in Hours". Save the table of data in a file because you will need it again in later exercises.

(e) Use the statistical package's plotting procedure to draw the scatterplot of the data, with the response variable shown on the vertical axis. (It is a convention that the response variable is shown on the vertical axis of graphs and plots because this convention helps viewers to quickly orient themselves to the displays.) Have the statistical package show the long variable names on the scatterplot since full naming of the axes on a graph or plot makes it easier to understand the display. Print a copy of the scatterplot so that you will be able to study it when you are away from the computer.

(f) Of course, each of the thirty points on the scatterplot represents one of the rows in the data table and also represents one of the entities in the research project. Draw small circles around the three points on the scatterplot that represent the first three rows in the table.

(g) Examine the scatterplot to see if any points on it are well away from the main cloud(s) of points. (Such points are called "outliers". Outliers may indicate errors in measurement, errors in transcription of the measurements, or unusual entities.) If you find any outliers that reflect typing errors that you have made, correct them and print the corrected scatterplot.

(h) From looking at the scatterplot, do you think a relationship exists in the population between the two variables? If so, describe the relationship.

Note that usually the type of education research project described in exercise 2 is done with at least one hundred students (and perhaps with several thousand students) and usually, to justify generalization, the students are selected from more than a single course, and perhaps from several different colleges or universities.

3. In exercise 2 you considered the relationship between the variables in the second and third columns in table 3. It is also conceivable that there is a relationship between the variables in the *first* and third columns of the table. That is, it is *conceivable* that there is a relationship between the student ID numbers and the grades of the students. However, although such a relationship is conceivable, most education researchers would be quite surprised if they were to find such a relationship because student ID numbers are usually arbitrary, and the arbitrariness usually makes such a relationship impossible. Although a relationship between student ID numbers and student grades is probably impossible, it is useful to examine the data for a relationship, to verify that the data are behaving as expected. (By looking at data from different points of view, one sometimes finds interesting new phenomena.)

   (a) Use the plotting procedure of the statistical package to draw the scatterplot of the possible relationship between student ID numbers and the student grades, with the grades shown on the vertical axis of the scatterplot. Have the statistical package show the long variable names on the scatterplot. Print a copy of the scatterplot so that you will be able to study it when you are away from the computer.

(b) Examine the scatterplot for outliers that reflect typing errors that you have made, and if you find any, correct them and print the corrected scatterplot.

(c) From looking at the scatterplot, do you think there is a relationship between student ID numbers and student grades? If so, describe the relationship.

4. A medical researcher wished to know if a proposed new blood pressure drug can control (i.e., lower) high blood pressure in people. Therefore, the researcher obtained a group of thirty patients who had registered with a particular hospital and who had high blood pressure. The researcher measured and recorded the average blood pressure of each patient. Next the researcher randomly divided the patients into two groups with fifteen patients in each group. Then, for the next ten weeks, using standard medical experimentation procedures, the researcher administered a daily dose of 100 milligrams of the new drug to the patients in one of the groups and a placebo (i.e., a pill with no medicinal ingredients) to the patients in the other group. At the end of the ten weeks, the researcher again measured the average blood pressure in each of the patients and determined the *drop* in average blood pressure in each patient. This yielded table 4.

   Note that some of the numbers in the rightmost column in the table are negative, indicating that the average blood pressure did not drop for these patients but instead rose. Note also that the predictor variable in this research project is "dose of the new blood pressure drug received". This variable has only two different values in the research project, namely, 0 milligrams per day and 100 milligrams per day. (Contrast this with the situation in exercise 2 in which the predictor variable has many different values.) Answer questions (a) through (h) in exercise 2 for the data in table 4. Note that often a more sophisticated research design is used in medical research than the simple design used in this exercise.

The research projects in exercises 2 and 4 demonstrate simple cases of the two main types of scientific research. The research project in exercise 2 is called an *observational research project.* This is because the researcher simply *observed* the response and predictor variables of interest. On the other hand, the research project in exercise 4 is called an *experiment* because, rather than merely observing the amounts of the blood pressure drug, the researcher actually "manipulated" the values of that variable by administering the different amounts of the drug to the patients. Experiments (i.e., research projects in which predictor variables are manipulated) are the main scientific method for obtaining information about causation and are the heart and soul of scientific research.

Table 4

Patient ID Number	Dose of New Drug (mg)	Drop in Average Blood Pressure (mm Hg)
1	100	-1
2	0	-20
3	100	33
4	0	-15
5	100	0
6	100	30
7	100	40
8	100	-13
9	0	-1
10	0	-19
11	100	20
12	100	12
13	100	8
14	0	33
15	100	39
16	100	9
17	0	-3
18	0	20
19	100	28
20	0	11
21	100	-6
22	100	35
23	0	10
24	0	-4
25	100	-21
26	0	-6
27	0	-36
28	0	-2
29	0	1
30	0	-1

7.10 A Definition of a Relationship Between Variables

(This subsection is included for completeness. Students who are uncomfortable with mathematical concepts can omit this material without losing the logical train of the paper.)

Let us consider a formal definition of a relationship between variables. We begin with some preliminary definitions and discussion:

Definition: A *numeric variable* is a variable whose values are numbers.

We can convert any variable that is not a numeric variable into a numeric variable by using a recoding rule to recode the non-numeric values of the variable into numbers. Thus, effectively, we can view all variables as

being numeric variables.

Definition: The *expected value* of a numeric variable in an entity in a population is the average value (i.e., the arithmetic mean value) of the variable across all the entities in the population under a given set of conditions.

Of course, the expected value of a variable is usually unknowable because it is usually not possible to measure the value of a variable in every entity in the associated population. However, we can *estimate* the expected value of any variable to any degree of precision we wish by measuring the value of the variable in each entity in an appropriate sample and then computing the average of the measured values. Such estimates are usually sufficient for our needs.

Definition: The *expected value* of a numeric variable in an entity in a population *conditioned on the values of one or more other variables* is the average value of the variable across all the entities in the population under a given set of conditions, including the condition that the other variables have particular stated values.

We shall represent the expected value of some variable y as $E(y)$, and we shall represent the expected value of y conditioned on the values of variables $x_1,\ldots,x_p$ as $E(y|x_1,\ldots,x_p)$.

Using the concept of expected value, let us now consider a formal definition of a relationship between variables:

Definition: If y is a variable that reflects a measured property of entities in some population, and if $x_1,\ldots,x_p$ are a set of one or more other variables that reflect distinct other measured properties of the entities (or of the entities' environment), then a *relationship* exists in the entities between y and the $x_1,\ldots,x_p$ if, for each integer i, where $1 \le i \le p$

$$E(y|x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_p) \neq E(y|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_p).$$

If $p = 1$, the inequality simplifies to

$$E(y|x_1) \neq E(y).$$

Each of the p inequalities is deemed to be satisfied if there is at least one set of specific values of the xs that satisfies the inequality. (Notes: (1) A different set of values of the xs may be used for each inequality; (2) if y is not a numeric variable, then for a relationship to exist, the p inequalities must each be satisfied for at least one recoding of the values of y into numeric values. A different recoding may be used for each inequality.)

This formal definition of a relationship between variables is operationally equivalent to the informal descriptions of a relationship between properties given at the end of subsection 7.1. That is, if a state of affairs satisfies the definition, it will also satisfy any of the descriptions, and vice versa.

In actual practice, scientists usually detect relationships between variables *not* by directly showing that the above definition is satisfied, but rather by showing that various other equivalent conditions are satisfied. These other conditions are designed to provide maximum efficiency in specific research situations and may at first appear to differ from the definition. However, satisfying these other conditions is equivalent to satisfying the definition in the sense that (if we ignore the ever-present, but controllable, "false alarm" errors) the other conditions will declare the existence of a relationship between variables only if the conditions of the definition are satisfied.

Other mathematical definitions of a relationship between variables or properties are given (in terms of causal relationships) by Bollen (1989, p. 41), Granger (1980), Chowdhury (1987), and Poirier (1988).

8. STATISTICAL TECHNIQUES FOR STUDYING RELATIONSHIPS BETWEEN VARIABLES

The preceding sections developed the concept of relationships between variables—a fundamental concept of science. In an introductory statistics course this concept leads us directly to a key question: What *statistical techniques* are available to help us study relationships between variables?

The available statistical techniques fall into four groups, namely,

- techniques for *detecting* relationships between variables
- techniques for *illustrating* relationships between variables
- techniques for *predicting* and *controlling* the values of variables, and
- miscellaneous techniques for studying relationships between variables.

Sections 9 through 13 discuss the four groups of techniques.

9. TECHNIQUES FOR DETECTING RELATIONSHIPS BETWEEN VARIABLES

One group of statistical techniques helps us to *detect* relationships between variables. Detecting relationships has been a source of much confusion and controversy.

9.1 Must Analyze Data

The only objective way of detecting a relationship between variables is to gather and analyze appropriate data that are likely, if a relationship exists, to reflect that relationship. We obtain such data by obtaining a table of

concurrent (or sometimes appropriately "lagged") measurements of the values of the variables of interest in a sample of entities from the population of interest. After we have obtained such a table, we can then use statistical methods (described below) to analyze the data in the table to look for evidence of a relationship.

9.2 The Null and Alternative Hypotheses

In detecting a relationship between a response variable and one or more predictor variables, a scientist will often partition the full set of possibilities into two simple opposing hypotheses:

- *Null Hypothesis:* there is *no* relationship between the response variable and any of the specified predictor variables in the entities in the population.
- *Alternative Hypothesis:* there *is* a relationship between the response variable and one or more of the specified predictor variables.

Of course, for any particular response variable and predictor variable(s) in any particular population of entities, one, and only one, of the preceding two hypotheses *must* be true.

For example, if we are interested in performing a research project to determine if a relationship exists between taking vitamin C and getting the common cold, the null hypothesis is that there is no relationship between vitamin C and the common cold. The alternative hypothesis is that there *is* a relationship between vitamin C and the common cold.

Formally, scientists begin the initial study of a relationship between variables with the assumption that the null hypothesis is true. Informally, however, scientists suspect and hope that some form of the alternative hypothesis is true. Scientists hope that the alternative hypothesis is true because if it is, then we (as society) can use information about the relationship to help make predictions or exercise control.

The practice of beginning with the (impossible-to-prove) assumption that the null hypothesis is true is entailed by the principle of parsimony, which tells us to keep things as simple as possible. The simplest situation is that of the absence of a relationship, so we begin with that assumption.

After making the formal assumption that the null hypothesis is true, scientists who wish to study a relationship then perform a research project in an attempt to invalidate the assumption. That is, as noted above, we obtain a table of the measured values of the relevant variables in the entities in a sample under circumstances in which the sought-after relationship between variables should appear, if it exists. We then analyze (as discussed below) the measured values in the table. If the analysis shows reasonable evidence that a relationship exists, then the scientific community (through informal consensus) rejects the null hypothesis and concludes that a relationship between variables similar to that suggested by the results probably does exist.

Neyman and Pearson (1928) introduced the concept of the null and alternative hypotheses. Fisher (1935, sec. 8) expanded these ideas and gave the null hypothesis its name.

Appendix A discusses two useful generalizations of the concept of a null hypothesis.

9.3 Why Should We Begin By Assuming That the Null Hypothesis Is True?

Tukey (1989, p. 176) suggests that a relationship may exist (albeit sometimes very weakly) in entities between *all* measurable pairs of variables, regardless of the identities of the variables. Given that suggestion, and despite the argument in subsection 9.2 in terms of the principle of parsimony, some students may wonder why it is necessary to begin a study of a relationship between variables with the assumption that the null hypothesis is true.

Scientists usually begin initial formal study of a relationship between variables with the assumption that the null hypothesis is true because this strategy prevents them from making the error of thinking that they know more about the relationship than they actually do. And only if all of the following conditions are satisfied will most scientists reject the null hypothesis and accept the existence of a particular relationship between variables:

- someone has performed an appropriate empirical research project
- the research project has found reasonable evidence of the relationship
- the research project has been carefully scrutinized for errors (and perhaps even successfully repeated) by other members of the scientific community (and anyone else who is interested), and
- nobody has come up with a reasonable alternative explanation of the results of the research project (Mosteller 1990, Lipsey 1990a, Macnaughton 1996b).

9.4 Statistical Tests for Detecting Relationships Between Variables

Various statistical methods are available to analyze the results of a research project and to help us determine if there is evidence of a relationship between the response variable and one or more of the predictor variables. These methods (which are usually computerized) work by taking as input the results of a research project. That is, the input is the table of values of the response variable and the predictor variable(s) that were measured in the entities that participated in the research project. Many of these methods provide, as output, one or more numbers called *p*-values.

If a research project and its data adequately satisfy the

underlying assumptions of a statistical method, then a *p*-value yielded by the method is an accurate estimate of

the fraction of the time
that we will obtain evidence
of the existence of the relationship
as strong as (or stronger than)
the evidence provided by the current research project,
if, actually, NO relationship exists
between the variables
in the population,
and if we were to perform the research project
over and over,
each time with a fresh random sample of entities
from the population.

The preceding paragraph implies that a *p*-value for a particular relationship will lie somewhere in the range between zero and one (including the two end-points).

The paragraph also implies that the lower a *p*-value, the less likely it is that the obtained evidence of the relationship would be obtained if no relationship exists. Thus if a *p*-value for a relationship is low enough, we can reject the null hypothesis and conclude that the associated relationship between the response variable and the indicated predictor variable(s) probably actually does exist in the population.

The *p* in *p*-value stands for *probability* because the fraction of the time that something happens is (one definition of) a probability.

How low should a *p*-value be before it is safe to conclude that the associated relationship exists? In evaluating the work of others, many scientists rely on the convention that a *p*-value should be lower than (or at least as low as) a "critical" value of .05 before they reject the null hypothesis and conclude that the relationship between variables that is associated with the *p*-value probably actually exists. (Some scientists use other critical *p*-values, notably .01.)

Thus the decision rule is simple: if a *p*-value computed from the results of the research project is less than (or equal to) the chosen critical *p*-value (and if the assumptions underlying the *p*-value are adequately satisfied), then we can (tentatively) reject the null hypothesis and conclude that the associated relationship exists in the population. But if the *p*-value is greater than the chosen critical *p*-value, all that we can conclude is that this particular *p*-value provides no conclusive evidence of the existence of the relationship.

The procedure of computing a *p*-value and examining it to determine whether it is less than a critical value is called a *statistical test* of the hypothesis that the associated relationship exists. Fisher (1925) and Neyman and Pearson (1928) began modern discussion of the concept of statistically testing a hypothesis. Lehmann (1986) gives a general discussion.

Consistent use of a critical *p*-value of .05 in statistical tests implies that if we consider all the times that we (properly) use statistical tests to determine if there is evidence of a relationship between variables, and if we then consider the portion of those times when the relationship suggested by the data does *not* actually exist, we will erroneously conclude that the relationship *does* exist in five percent of those times.

Many different statistical tests are available to help us detect relationships between variables, with the choice of the optimal test depending on the situation at hand. These tests all work in the general manner described in this subsection.

Careful readers need enough detail of the theory of statistical tests to allow them to decide whether to believe that statistical tests are valid. In particular, careful readers need to know details about the computation of *p*-values so that they can judge whether *p*-values are reasonable estimates of the probabilities they purport to estimate. These readers may find it helpful to note that each *p*-value is based on a "test statistic", which is simply a number computed from the results of a research project by following a well-defined mathematical procedure. Many statistics textbooks describe the procedures to compute various test statistics from the results of an appropriate research project. Statistical theory shows that, if the null hypothesis is true (and if certain other often-satisfiable underlying assumptions are adequately satisfied), then the associated test statistic will have a known "distribution". (The distribution is simply a description of what the different values of the test statistic will be with what frequency, if the test statistic were to be computed over and over, each time with a fresh random sample of entities from the population of interest.) But if the null hypothesis is false, the test statistic is specifically designed so that its distribution will usually deviate as far as possible from the distribution that occurs when the null hypothesis is true. Computation of a *p*-value is simply computation of the probability (i.e., fraction of the time) that the given test statistic will, if the null hypothesis is true, deviate as far as it has from its expected value under the distribution of values that is known to occur when the null hypothesis is true. Many statistics textbooks describe the procedures to compute these probabilities.

Readers who wish to know more about the mathematical details of statistical methods will find helpful discussion in introductory statistics textbooks. Recommended are the books by Moore (1995), Iman (1994), Freedman, Pisani, Purves, and Adhikari (1991), and Snedecor and Cochran (1989). Discussion also appears in most of the works listed at the end of subsection 11.1.

Creative readers new to the material may rightfully wonder whether there might be a general objective procedure for detecting relationships between variables that is

easier to use than the *p*-value approach. Unfortunately, an acceptable easier procedure has not yet been invented, and perhaps no such easier procedure is possible.

The use of a statistical test provides a clear-cut criterion for deciding when to believe a relationship exists between variables. Another approach to detecting relationships between variables is to derive an estimate of the *probability* that a particular relationship exists. However, a problem with deriving the probability that a relationship exists is that such a probability is associated with *one particular form* of the relationship. Instead of being tied to a particular form of the relationship, scientists often prefer a more general criterion of determining whether *any type* of relationship exists. The *p*-value generally comes close to providing such a criterion, and thus the *p*-value is generally preferred.

It is important to note that a low *p*-value does not imply that the associated relationship between variables has any *practical* implications. Some relationships between variables, when discovered, will have immediate practical implications, but other relationships will not. It is also important to note that, even if a new relationship has no obvious practical implications, it should generally still be reported in the scientific literature because knowledge of the relationship may stimulate other researchers to look further into the relationship, and practical implications may then be found.

Rather than consistently using the same fixed critical *p*-value (e.g., .05), some authors have suggested that we can determine the critical *p*-value for an instance of a particular statistical test on the basis of one or more aspects of the particular research situation in which the test appears. However, as a practical matter, it is often difficult to find *objective* aspects of a research situation that enable us (perhaps through a "loss function") to determine an appropriate critical *p*-value for a statistical test in that situation. On the other hand, many scientists feel that the use of *subjective* criteria to determine a critical *p*-value is arbitrary, and thus less desirable in science, which tries to be as objective as possible. Thus we are often led to using fixed critical *p*-values, such as .05 or .01. The use of fixed critical *p*-values is also reinforced by some editors of scientific journals who require that the important *p*-values in a research paper be less than a fixed value (often .01) before they will consider the paper for publication. This requirement helps editors control the frequency of publishing "false alarms", which are discussed in the next subsection.

9.5 The Four Possible Outcomes of a Statistical Test

If we use a statistical test to determine whether to conclude that a relationship exists between variables in a population of entities, there are four possible outcomes of the test that can occur. The four outcomes are shown in table 5.

Table 5. The Four Possible Outcomes of a Statistical Test of a Relationship Between Variables

	REALITY IS	
THE STATISTICAL TEST SAYS ↓	*The relationship exists*	*The relationship does not exist*
"Conclusive evidence of a relationship was found." i.e., $p \leq p_{crit}$	OUTCOME: correct detection of the relationship	OUTCOME: false alarm error
"Conclusive evidence of a relationship was not found." i.e., $p > p_{crit}$	OUTCOME: miss error	OUTCOME: correct failure to detect any relationship

The rightmost two columns of the table represent the two possibilities with respect to the existence in the population of the studied relationship between variables. That is, either the relationship exists (middle column of the table) or the relationship does not exist (rightmost column of the table).

Similarly, the bottom two rows in the table represent the two possibilities with respect to the statistical test. That is, either the *p*-value is less than (or equal to) the critical *p*-value (which implies that the statistical test has found "conclusive" evidence that a relationship exists) or the *p*-value is greater than the critical *p*-value (which implies that the statistical test has not found conclusive evidence that a relationship exists).

The goal of performing a statistical test is to determine which of the rightmost two columns of the table is correct for the particular relationship between variables under study. We make this determination by noting in which of the bottom two rows of the table the statistical test places us. If we find ourselves in the second row from the bottom, then we reject the null hypothesis and (tentatively) conclude that a relationship exists. But if we find ourselves in the bottom row of the table, then we conclude that we do not have sufficient evidence to conclude that a relationship exists.

The four cells in the table labeled OUTCOME constitute the "body" of the table. These cells represent the four possible outcomes of a statistical test. Let us examine each outcome in turn.

The upper-left cell in the body of the table represents the outcome in which the suggested relationship actually exists between the variables in the population and the statistical test provides conclusive evidence that the relationship exists. This is the outcome that we would always like to occur, because whenever it does occur we have found a real relationship between variables in the population.

The lower-right cell in the body of the table represents the outcome in which the suggested relationship between variables does *not* exist in the population and the statistical test correctly indicates the absence of conclusive evidence of a relationship (i.e., the *p*-value is greater than the critical *p*-value). This outcome is less desirable than the outcome described in the preceding paragraph, because it implies that a relationship that we had hoped to find does not actually exist. However, this outcome is more desirable that the remaining two outcomes because these outcomes both represent errors.

The upper-right cell in the body of the table represents the outcome in which the suggested relationship between variables does not exist in the population, but the statistical test erroneously indicates that the relationship does exist. Using terminology from signal detection theory, let us call this type of error a *false alarm*, because it amounts to our falsely concluding that a relationship exists. In the preceding subsection we noted that the critical *p*-value indicates the fraction of the time that we will mistakenly conclude that a relationship between variables exists in the population, when such a relationship does not actually exist. Thus the critical *p*-value can be characterized as the false alarm error rate.

(False alarm errors are also sometimes called type I [i.e., type one] errors.)

The lower-left cell in the body of the table represents the outcome in which the suggested relationship exists between the variables in the population, but the statistical test does not succeed in detecting the relationship. Using terminology from signal detection theory, let us call this type of error a *miss* error, because it amounts to our missing detecting that the relationship exists. (Miss errors are also sometimes called type II [type two] errors.)

Some beginners mistakenly believe that use of a critical *p*-value of, say, .05 implies that we will commit false alarm errors five percent of the times that we use statistical tests. That is, some beginners mistakenly believe that we will end up in the upper-right cell in the body of table 5 five percent of the times that we use statistical tests. The correct interpretation is that a critical *p*-value of .05 implies that we will commit false alarm errors five percent of the times *that the indicated relationship does not exist.* In other words, in research situations in which reality places us in the rightmost column of the table, a statistical test with a critical *p*-value of .05 will, at random, send us to the upper-right cell in the body of the table in five percent of those situations, and it will send us to the lower-right cell in the body of the table in the other ninety-five percent of those situations.

Having considered the relative frequency with which statistical tests will send us into the two rows in the rightmost column of the body of the table, let us now consider the relative frequency with which we will fall into the two columns in the body of the table. Here, it is important to note that many (but not all) researchers study relationships between variables that they have good reason to believe to exist. (This is because to study relationships believed *not* to exist has little or no payoff. Studying relationships believed not to exist amounts to trying to support the null hypothesis, which most scientists simply *assume* to be true until satisfactory evidence to the contrary is brought forward.) Therefore, because researchers have usually carefully chosen the relationships they study, they will often (but certainly not always) fall in the middle column of the table.

Because the use of statistical tests to detect relationships between variables implies that (1) we will sometimes make false alarm errors, and (2) we will sometimes make miss errors, it is important to consider the consequences of both types of errors.

In the case of false alarm errors it is gratifying to observe that, if scientists mistakenly conclude that a certain relationship exists between variables when, in fact, there is no such relationship, then the mistake is usually rapidly detected through the process of replication—i.e., independent demonstration—of the relationship. Whenever a research project reports an interesting new relationship between variables, other researchers usually attempt to replicate the relationship. If other researchers are consistently unable to replicate a relationship between variables, the existence of the relationship is called into question, and the scientific community may return to assuming that the null hypothesis is true.

For example, in the recent "cold fusion" controversy, it is helpful to view the controversy as a question of whether certain newly claimed relationships between variables are present in a "cold fusion cell". In particular, some scientists have reported that, under certain conditions, the heat energy output (response variable) of a cold fusion cell is greater than the electrical energy input (predictor variable), which is contrary to accepted physical theory. However, other scientists have been consistently unable to replicate this relationship. Because many scientists have been unable to replicate the new relationship, most scientists now assume that the claimed new relationship was a false alarm, and the relationship does not exist. Therefore, cold fusion is considered, at this time, not to be possible (Huizenga, 1993).

In contrast to false alarm errors, the consequences of miss errors can be more damaging. First, much of scien-

tific research is done directly or indirectly to search for benefits for humankind. When scientists or statistical tests commit a miss error (i.e., they fail to detect a real relationship between variables in a population), study of the relationship may be abandoned, and therefore, implementation of useful applications of the relationship may be delayed, perhaps with a serious social cost. Second, on a more personal level, if a scientist commits a miss error and fails to find a useful new extant relationship between variables, then he or she will also miss receiving the various rewards associated with finding the relationship. Therefore, scientists usually take steps to ensure that the "miss error rate" for important statistical tests in their research projects is appropriately low. These steps are introduced in the next subsection.

9.6 The Power of Statistical Tests For Detecting Relationships Between Variables

Definition: The *power* of a statistical test for detecting a particular relationship between variables is the *fraction of the times* that the test is performed (each time with a fresh random sample of entities from the population of interest) *that the test will detect the relationship*, given that the relationship has a particular form.

The above definition implies that the power of a particular statistical test for detecting a particular relationship is a particular number lying somewhere in the range between zero and one (including the two endpoints).

As we might expect, the power of a particular statistical test for detecting a particular relationship depends on various properties of the research project in which the test appears. For example, the more accurate the methods we use to measure the values of the relevant variables in the entities in a research project, the more powerful the statistical tests. Similarly, the more entities that (properly) participate in a research project, the more powerful the statistical tests.

Of course, when we use statistical tests to help us detect relationships between variables, we are usually eager to discover new relationships. Therefore, we usually want to use statistical tests that have (when operating in the vicinity of the expected form of the relationship) the highest possible power. Therefore, scientists usually design research projects so that the power of the important statistical tests (for detecting the expected form of the sought-after relationship) is at least .8 and sometimes as high as .99.

To help scientists design research projects, statisticians have developed methods to compute the power of most of the statistical tests that are used for detecting relationships between variables. These methods allow a researcher to verify, ahead of time, that the statistical tests he or she plans to use in a research project will have sufficient power. Some of these methods are discussed by Odeh and Fox (1991), Lipsey (1990b), and Kraemer and Thiemann (1987). Also, some statistical software packages now contain programs that automatically compute the power of statistical tests.

Since the power of a statistical test is the fraction of the times that when the test is performed it *will* detect the stated relationship, therefore one minus the power is the fraction of the times that when the test is performed it *will not* detect the stated relationship. Thus one minus the power is the *miss error rate* of the statistical test used to detect the relationship.

The miss error rate is the fraction of the time that, when we are in the *middle* column of table 5, the statistical test will erroneously place us in the *bottom* row, given that the relationship has the stated form. Contrast the miss error rate with the false alarm error rate (which we noted in the previous subsection is equal to the critical *p*-value). The false alarm error rate is the fraction of the time that, when we are in the *rightmost* column of table 5, the statistical test will erroneously place us in the *top* row of the body of the table.

Appendix B lists factors that we can control in a research project to maximize the power (i.e., minimize the miss error rate) of statistical tests that detect relationships between variables.

Most research projects that use statistical tests to detect relationships carry out no more than ten statistical tests that are crucial for the aims of the research project. (In fact, many research projects carry out only a single crucial test.) However, a few research projects carry out more than ten crucial tests, perhaps many more. Appendix C discusses power considerations for research projects that carry out many crucial statistical tests.

9.7 Why Do We Need Statistical Tests?

Of course, we need not use a statistical test if we discover a very strong relationship between variables, because in this case, the results of the research project usually leave no doubt about the existence of the relationship. However, new strong relationships between variables are not often discovered, perhaps because most of the strong relationships have already been discovered. Thus most relationships that are currently studied are weak enough that statistical tests are necessary.

We noted in subsection 9.3 that Tukey (1989, p. 176) suggests the existence of a relationship (albeit sometimes very weak) in entities between *all* measurable pairs of variables, regardless of their identities. Given Tukey's suggestion, it is reasonable to ask why we *need* to use statistical tests to demonstrate evidence of a relationship between a response variable and a selected set of predictor variables when, in fact, almost surely, there is a relation-

ship.

We need statistical tests because, as Tukey notes, in addition to wishing to know with confidence whether a relationship exists, we usually also need reliable information about the "profile" of the relationship.

The *profile* of a relationship is effectively the shape and orientation of the line that shows the relationship on a graph. The profile tells us whether, when the value of a given predictor variable increases in entities, the value of the response variable can be expected to *increase* or to *decrease*, or perhaps to sometimes increase and sometimes decrease, depending on the values of other variables.

We need statistical tests because they allow us to achieve an objective level of confidence that the modest relationships between variables that are typically suggested by the results of modern research accurately reflect profiles we can expect to find in other entities in the population. Only if we have accurate profiles can we hope to make accurate predictions or exercise accurate control.

9.8 Graphical Techniques For Detecting Relationships Between Variables

In addition to using statistical tests, we can also (or instead) use graphical techniques to help us detect relationships between variables. Graphical techniques rely on the highly developed ability of the human eye to detect patterns in visual stimuli.

Graphical detection techniques involve the joint presentation of the values (or functions of the values) of (*a*) the response variable, and (*b*) one or more of the predictor variables for the entities in a research project in one or more graphical displays. Popular types of graphical displays include graphs, scatterplots, boxplots, and bar charts. (Most modern statistical software packages can generate extensive graphical displays upon the simple request of a user.) After the appropriate graphical displays are obtained, we carefully search them for patterns that suggest a relationship in the population between the response variable and the displayed predictor variables.

Cleveland (1993) and Tukey (1988) discuss graphical techniques for detecting relationships between variables.

9.9 Comparison of Numerical and Graphical Techniques For Detecting Relationships Between Variables

Theoretically, any relationship between variables that can be detected with graphical detection techniques can also be detected with numerical detection techniques, and vice versa. However, experienced researchers have failed to detect certain relationships with numerical techniques that experienced researchers have later detected using graphical techniques, and vice versa.

If a researcher fails to detect a relationship that is clearly present in data, then the failure occurred because the researcher did not look (numerically or graphically) carefully enough at the data. Researchers sometimes fail to look carefully enough at their data because there are currently no complete algorithms that researchers can follow when trying to find relationships between variables. (A complete algorithm must include full "residual" and "outlier" analyses—Anscombe and Tukey 1963, and Barnett and Lewis 1994.) Thus researchers are left to their own devices, and they sometimes fail to detect useful visible relationships. To alleviate this problem, statisticians will probably develop complete formal relationship-detection algorithms. After such algorithms have been computerized and validated (by testing and refining them on a large sample of real data sets), fewer relationships, or aspects of relationships, will be overlooked.

Some advantages and disadvantages of numerical and graphical techniques for detecting relationships are as follows:

- graphical techniques are usually easier to understand
- numerical techniques are necessary if we need an objective measure of the believability that a suspected relationship exists (e.g., the *p*-value)
- numerical techniques (properly applied) are more likely than graphical techniques to detect a relationship if the relationship is extremely weak or if the relationship is subtle and involves many predictor variables.

Because both numerical and graphical detection techniques have advantages, researchers sometimes use a combination of the two approaches.

9.10 Exercises

1. Refer to the scatterplot that you printed for exercise 2 in subsection 7.9.
   (a) Before you attempt the remaining parts of this exercise, reconsider the scatterplot and your conclusion about whether you think a relationship exists between the two variables. Have you changed your mind?
   (b) For the possible relationship between the two variables shown on the scatterplot, what are the null and alternative hypotheses?
   (c) Use the linear regression procedure in your statistical package to analyze the data underlying the scatterplot and compute the *p*-value for a linear relationship between the predictor variable and the response variable.
   NOTE: Your instructor will provide you with instructions for using the linear regression procedure or will provide you with a reference to instructions in the user manual for the statistical package. To get correct results you must correctly specify to the procedure which variable is the response variable and which is the predictor variable. Unless you have an unusual statistical package, you should not

have to specify any options to the procedure because the *p*-value and the other statistics you will need should be provided in the default output. (You can ignore all of the other statistics at this time.) Statistical packages use various labels for the *p*-value in their documentation and output, such as p, P, P(2 TAIL), Prob>F, Prob > |T|, Signif F, and Sig T. Also, most regression procedures will give more than one *p*-value for your data. The second *p*-value is related to the "intercept". If a *p*-value is lower than around .0001, some statistical packages show the *p*-value as zero, even though the true *p*-value is always greater than zero. Other statistical packages show low *p*-values using E-notation. For example, a *p*-value of .000000002 could be shown as 2.E-09.

NOTE: The use of linear regression to compute a *p*-value is based on certain assumptions. The assumptions are that (1) a straight line—rather than a curved line—is appropriate for the points on the associated scatterplot, and (2) if lines are drawn *vertically* from each of the points to the best-fitting line for the points on the scatterplot, the signed lengths of these vertical lines will have an "independent normal distribution" with a constant "variance" across the different values of the predictor variable. Details of these assumptions are beyond the scope of this paper. It is, however, very important to note that, to avoid later possible embarrassment, one should always verify that the underlying assumptions of a statistical procedure are adequately satisfied before attempting to draw conclusions from use of the procedure in a real-life research project.

(d) Assume that you have verified that the underlying assumptions are adequately satisfied for the *p*-value you obtained in part (c). Given that fact, and given the *p*-value you have obtained, is it reasonable to conclude that a relationship between the two variables exists in the population?

NOTE: For the "homework - grade" example, the underlying assumptions are clearly violated because the vertical distributions of the points about the best-fitting line are "truncated". That is, if this is a typical course, the values of the students' grades cannot go above 100 or below zero. (A normal distribution is *not* truncated, so the points cannot lie in a normal distribution, so the assumptions are violated.) However, the *p*-value in linear regression is "robust" to this minor violation of the assumptions. Furthermore, in this case violation of the assumptions does not lead us to doubt the existence of the relationship because the relationship between the two variables is strong enough to allow one to easily see the presence of the relationship by merely looking at the scatterplot, without having to rely on the *p*-value. If the relationship were weaker, it would be reasonable to analyze the data with a procedure whose underlying assumptions would not be violated by the truncated distributions in order to corroborate the existence of the relationship. (Researchers generally use such alternative procedures only when necessary because the statistical tests in such procedures are slightly less powerful than the statistical tests in multiple regression.)

2. Refer to the scatterplot that you printed for exercise 3 in subsection 7.9. Answer questions (a) through (d) in exercise 1 for the data underlying that scatterplot.
3. Refer to the scatterplot that you printed for exercise 4 in subsection 7.9. Answer questions (a) through (d) in exercise 1 for the data underlying that scatterplot.

   NOTE: The use of linear regression to obtain the *p*-value for the data in exercise 3 is somewhat unusual, and many researchers would instead use a "two-sample *t*-test" to obtain the *p*-value. Other researchers would use a "one-way analysis of variance" (one-way ANOVA, a generalization of the *t*-test) to obtain the *p*-value. However, in the case of a single predictor variable that has only two different values in the research project, it can be shown that the *p*-value obtained with the linear regression is identical to both that obtained with the two-sample *t*-test and that obtained with one-way ANOVA. (The *t*-test and the associated one-way ANOVA have the same underlying assumptions as the linear regression except that the assumption about a straight line is no longer relevant.) The use of linear regression to compute the *p*-value illustrates how the same underlying procedure is used to compute *p*-values for both observational data and experimental data.
4. Use the two-sample *t*-test procedure in the statistical package to compute the *p*-value for the relationship between the amount of the drug and the drop in blood pressure in table 4 in subsection 7.9. This *p*-value should be identical to the *p*-value you obtained in exercise 3.
5. Use the ANOVA procedure in the statistical package to compute the *p*-value for the relationship between the amount of the drug and the drop in blood pressure table 4 in subsection 7.9. This *p*-value should be identical to the *p*-value you obtained in exercise 3.

10. TECHNIQUES FOR ILLUSTRATING RELATIONSHIPS BETWEEN VARIABLES

The preceding section discussed how statistical techniques can help us to *detect* relationships between variables. A second group of statistical techniques helps us to *illustrate* these relationships. Illustrating a relationship

usually helps us to understand it. Also, illustrating a relationship is usually the most effective way of communicating information about the relationship to others.

Most of the techniques for illustrating relationships have been computerized, and thus we use them by feeding the data table from the sample and some simple instructions to a computer, whereupon the computer generates the appropriate graphical display(s) to illustrate the specified relationship in the data.

As noted in subsection 7.9, it is a convention that if the values of the response variable in a research project are shown in a graphical display, these values are shown on the *vertical* axis of the display.

Cleveland (1993) and Tukey (1988) discuss techniques for illustrating relationships. Tufte (1983, 1990) gives a general discussion of the use of graphical displays to convey information, especially information about relationships between variables.

11. TECHNIQUES FOR PREDICTING AND CONTROLLING THE VALUES OF VARIABLES

The preceding two sections discussed how statistical techniques help us to *detect* and *illustrate* relationships between variables. A third group of statistical techniques helps us to satisfy a central goal of science, which is to *predict* and *control* the values of variables, as was discussed in section 6. The present section focuses on prediction techniques.

(Control techniques are similar to prediction techniques, but the require "manipulation" of the predictor variables in the research project to allow us to determine whether the relationships we discover are "causal" relationships. Only if we confirm that relationships are causal are we ensured control capability. Detailed discussion of control techniques is beyond the scope of this paper, but see Macnaughton [1996b, chap. 4].)

Of course, when we make predictions, we wish to predict the values of a particular response variable in new entities in the population of interest. We make the predictions on the basis of the values of one or more predictor variables in the new entities and on the basis of our knowledge of the relationship between the response variable and the predictor variable(s).

Statistical predictions are usually not perfectly accurate. This lack of perfect accuracy is due to our current incomplete knowledge about most relationships and our current less-than-perfect measurement methods. However, as we shall see below, statistical prediction techniques are designed to give the "best" predictions possible.

11.1 Prediction Techniques

If we discover that a relationship exists between variables, and if we have used different groups of entities in the original research project (and if none of the predictor variables participate in the analysis as "continuous" variables), then we can make predictions of the value of the response variable for new entities by using the average values of the response variable for the various groups. (A variable is "continuous" if it can theoretically have any value within some continuous range of numerical values.) That is, usually the best prediction of the value of the response variable for an entity is the average value of the response variable for the group of all the entities in the original research project that had values of the predictor variable(s) that are highly similar to the value(s) that the new entity has. We can read such predicted values from a table of average values or from a graphical display of the information in the table.

The approach described in the preceding paragraph works when appropriate groups were used in a research project and the group average values of the response variable are available (and meaningful). In other cases, scientists often make predictions by means of a *model equation* (sometimes simply called a *model*). A model equation is a mathematical equation that expresses the predicted value of a response variable in terms of some mathematical function of the values of one or more predictor variables.

For example, a simple model equation is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2,$$

where $\hat{y}$ is the predicted value of the response variable for a new entity from the population, and x_1 and x_2 are the values of two predictor variables for the entity.

(The hat [ˆ] on the y in the equation indicates that it is a predicted value, in order to distinguish it from the "true" measured value.)

The b_0, b_1, and b_2 in the equation are called the *parameters* of the equation and are presumed to represent fixed numbers. Most model equations have one or more parameters. Parameters may reflect constants of proportionality or other numbers (such as the exponent 2 in $E = mc^2$). Of course, before we can use a model equation for making predictions, we must first obtain estimates of the values of all the parameters in the equation.

Various statistical methods are available to help us determine the form of the "best" model equation for expressing a relationship between variables and to help us obtain estimates of the values of the parameters in that equation. These methods (which are often computerized) take as input the table of values of the response variable and the predictor variable(s) that were measured in the entities that participated in a research project. The methods provide, as output, a statement of various candidate forms of the model equation and estimates of the values of the parameters in those equations.

Scientists usually choose the best model equation from among the candidates by balancing two opposing re-

quirements:

- choose the form of the model equation and the estimates of the values of the parameters so that the equation gives the smallest errors in prediction ("residuals") if it is used to *predict* the values of the response variable for all the values of the response variable that were *actually obtained* in the research project
- choose the form of the model equation so that it has as few parameters as possible (in order to satisfy the principle of parsimony).

Statistical methods for determining model equations and statistical methods for estimating the values of parameters provide various techniques to help us balance these two requirements so that we can obtain the "best" model equation.

Appendix D discusses two popular ways of choosing the values of parameters in a frequently used type of model equation so as to minimize the prediction errors.

Once we have determined (*a*) the form of a model equation for a relationship between variables and (*b*) estimates of the values of the parameters in that equation, we can then easily make predictions for new entities by substituting into the equation the value(s) of the predictor variable(s) for a new entity. After making the substitution, we can then do the required arithmetic to obtain the predicted value of the response variable for the entity.

In some cases, especially in the hard sciences, model equations are not derived from the results of a research project, but are instead derived from theoretical considerations. For example, Einstein used theoretical considerations to derive his model equation $E = mc^2$. In such cases we need not use statistical methods to *derive* a model equation from the results of a research project because the equation will already have been derived. However, we can still use statistical methods (especially residual analysis) to examine the results of a research project that studies the relationship implied by a theoretically derived model equation to help us determine whether the equation is consistent with the external world. This helps us to confirm that the theoretical derivation of the equation is correct.

The powerful numerical and graphical methods that scientists use to help them determine the form of a model equation and obtain estimates of the values of the parameters include the following:

- experimental methods discussed by Myers and Montgomery (1995), Mason, Gunst, and Hess (1989), Box and Draper (1987), Box, Hunter, and Hunter (1978), Winer (1971), Tukey (1992), Cox (1958), Cochran and Cox (1957), Anderson and Bancroft (1952), Kempthorne (1952), and Fisher (1935, 1925), with important theoretical discussions by Searle (1987), Hocking (1985), Graybill (1976), Rao (1973), and Scheffé (1959)
- multiple regression methods discussed by Montgomery and Peck (1992), Chatterjee and Price (1991), Kleinbaum, Kupper, and Muller (1988), Weisberg (1985), and Draper and Smith (1981)
- survey methods discussed by Yates (1981), Cochran (1977), Kish (1965), Hansen, Hurwitz, and Madow (1953), and Deming (1950)
- methods of the generalized linear model (and derivative methods) discussed by McCullagh and Nelder (1989) and Nelder and Wedderburn (1972)
- robust and exploratory methods discussed by Hoaglin, Mosteller, and Tukey (1991, 1985, 1983), Mosteller and Tukey (1977), and Tukey (1977)
- categorical methods discussed by Agresti (1990), Koch, Carr, Amara, Stokes, and Uryniak (1990), Cox and Snell (1989), Goodman (1984), Plackett (1981), Fienberg (1980), Goodman and Kruskal (1979), Haberman (1978, 1979), Bishop, Fienberg, and Holland (1975), and Mosteller (1968)
- time series methods discussed by Kendall and Ord (1990), Anderson (1971), Box and Jenkins (1970), and Tukey (1985, 1984)
- Bayesian methods discussed by Bernardo and Smith (1994), Howson and Urbach (1993), Good (1992), Lee (1989), Press (1989), Cyert and DeGroot (1987), Dawid (1986), Berger (1985), Mosteller and Wallace (1984), Hartigan (1983), Lindley (1983), Box and Tiao (1973), Winkler (1972), and Zellner (1971)
- linear structural relationship methods discussed by Jöreskog (1993) and Bollen (1989).

In addition to discussing deriving model equations, most of the works in the preceding list also discuss methods of detecting and illustrating relationships between variables.

Appendix E lists factors that we can control in designing a research project in order to maximize the accuracy of later predictions or control of the values of the response variable for new entities from the population.

11.2 A Measure of Prediction Accuracy

Researchers sometimes report the accuracy of a prediction by means of a *confidence interval.* A confidence interval consists of two pieces of information: (*a*) an interval (i.e., range) of values around the predicted value of the response variable and (*b*) a percentage.

A confidence interval for a predicted value of a response variable tells us (provided that certain underlying assumptions are adequately satisfied) that if the actual measured values of the response variable are (after they become available) compared with the associated predicted values, the measured values will lie within the ranges (about the predicted values) specified by the associated confidence intervals the indicated percentage of the times that predictions are made.

For example, a researcher might determine and report that the actual measured values of a particular response

variable will lie within 2.8 units of the values predicted by a particular model equation ninety-five percent of the times that the equation is used to make predictions.

Confidence intervals are useful indicators of prediction accuracy. However, they should be used with caution because it is often difficult to tell whether a prediction situation is similar enough to the original research situation for a given confidence interval to be accurate.

11.3 Prediction With No Predictor Variables

When studying statistical techniques for making predictions, we can include the empirical study of single variables within the concept of a relationship between variables. We can do this by considering the study of the different values of a single variable in the entities in a population of entities that was studied in a research project to be simply a special case of a relationship between variables. This special case is a *degenerate* case in which the response variable is present, but the set of predictor variables is empty (i.e., has no members). This point of view is consistent at a deep level with the underlying mathematics of the study of relationships between variables. (It is consistent in the sense that the underlying mathematics of the no-predictor-variable case is the limiting case of the mathematics of a relationship between variables when the number of predictor variables is reduced to zero.)

Cases of prediction with no predictor variables occur infrequently in scientific research because we are assured better ability to predict or control a response variable if we study the relationship between that variable and relevant predictor variables, rather than if we simply study the response variable in isolation.

11.4 Exercises

1. Refer to the scatterplot that you printed for exercise 2 in subsection 7.9. Scanning this scatterplot suggests that a straight line can be drawn through the cloud of points on the scatterplot to summarize the relationship between the two variables.
   (a) Use a ruler to draw on the scatterplot what appears to be the best straight line through the points to summarize the relationship between the students' grades and the number of hours of homework they did.
   (b) The model equation for a straight-line relationship between two variables has the form
   $$\hat{y} = b_0 + b_1 x$$
   where
   $\hat{y}$ = the predicted value of the response variable for an entity
   x = the value of the predictor variable for the same entity
   b_0 = a parameter of the model equation (called the "intercept" of the best-fitting straight line)
   b_1 = a second parameter of the model equation (called the "slope" of the best-fitting straight line).
   Use the linear regression program in your statistical package to analyze the data underlying the scatterplot and find the estimates for b_0 and b_1 in the output from the program. The default line that linear regression programs use to model the relationship between two variables is a straight line. Therefore, it should not be necessary to specify any options to the program because the estimates of b_0 and b_1 should be present in the default output. Different regression programs use different names for these parameters. The parameter b_0 is sometimes called "b0", "B0", "beta0", "constant", and "intercept". The parameter b_1 is sometimes called "b1", "B1", "beta1", "coefficient", the actual name of the predictor variable, and "slope".
   (c) Write the model equation for the relationship between the two variables as it is written above in part (b) except use the actual numerical values of b_0 and b_1 in the equation [as you obtained in part (b)] instead of b_0 and b_1.
   (d) Use the equation you wrote in part (c) to predict the grade for a student in the population who does 10 hours of homework in the course. Use the equation to predict the grade for a student who does 65 hours of homework, and for a student who does 135 hours of homework.
   (e) Plot points for the three predicted values you obtained in part (d) on the scatterplot. You can use the following procedure to plot a point for a predicted value on a scatterplot:
      i. Locate the value of the predictor variable on the horizontal axis of the scatterplot. Call this point A.
      ii. Draw a light pencil line vertically from point A.
      iii. Locate the value of the response variable on the vertical axis. Call this point B.
      iv. Draw a light pencil line horizontally from point B to intersect the vertical line you drew from point A. The point of intersection is the point on the scatterplot for the given values of the predictor and response variables.
   Because the prediction equation is a linear (i.e., straight line) equation, the three points you plotted should lie on a straight line. If the points do not lie on a straight line, find the error(s) and replot the points so that they lie on the correct straight line.
   (f) Join the three points that you plotted in part (e) with a straight line. This is the best-fitting line for

the scatterplot as provided by linear regression, and is called the "least-squares" line. (This line is defined in appendix D). Compare the least-squares line with the line you fitted by eye in part (a).

(g) Some statistical packages can automatically draw the least-squares straight line on a scatterplot. If your package can do so, have it redraw the scatterplot with the least-squares straight line shown on the scatterplot.

(h) Use the least-squares line on the scatterplot to predict the value of the course grade for a new student in the population if that student does 100 hours of homework. You can use the following procedure to make the prediction:

i. Locate the point on the horizontal axis the represents the value of the hours of homework done by the student. Call this point A.

ii. Draw a vertical line from point A to intersect the best-fitting line on the scatterplot and call the point of intersection point B.

iii. Draw a horizontal line from point B to intersect the vertical axis and call this point of intersection point C. The predicted value of the course grade for the student is equal to value represented by point C on the vertical axis of the scatterplot.

(i) Find the statistic in the computer output labeled the "standard error (of the predictions)". (Some packages call this statistic the "standard error of the estimates", the "Root MSE" [square root of the mean squared error] or "s".) The standard error of the predictions is measured in the same units as the predictor variable. In observational research projects, if certain assumptions are satisfied, then twice the standard error of the predictions slightly underestimates the size of a 68 percent confidence interval (centered on the predicted value) for individual predictions of the value of the response variable. (The assumptions are that [1] the predictions are made for members of the same population who participated in the research project, [2] the relevant underlying assumptions of the statistical procedure [i.e., linear regression] are adequately satisfied, [3] a "sufficient" number of entities participated in the research project, and [4] all the relevant conditions in the prediction situation are the same as the conditions of the research project.) That is, if the least-squares line is used to make predictions of grades for new students (and if all the assumptions are satisfied), then we can expect slightly fewer than 68 percent of the predicted grades to be no more than the size of the standard error away from (above or below) the corresponding true measured course grades for the students. The standard error of the predictions is a rudimentary method for estimating prediction accuracy. More sophisticated methods are described by Hahn and Meeker (1991).

2. Refer to the scatterplot that you printed for exercise 4 in subsection 7.9. A straight line can be drawn on this scatterplot through the centers of the two vertical clouds of points in order to summarize the relationship between the two variables.

(a) Use a ruler to draw on the scatterplot what appears to be the best straight line through the points to summarize the relationship between the amount of the drug received and the drop in blood pressure.

NOTE: It may be presumptuous to assume that a straight line is appropriate for the relationship because we have no data *between* the two values of the predictor variable. (For example, we have no data for values of the drop in blood pressure when patients receive 50 milligrams of the drug each day.) Nevertheless, since a straight line is the simplest case, we shall invoke the principle of parsimony and assume that a straight line is appropriate until evidence to the contrary is brought forward in other research.

(b) If we assume that a straight line is appropriate, we can use the model equation as described in part (b) of exercise 1 to summarize the relationship. Use the linear regression program in your statistical package to analyze the data underlying the scatterplot and find the estimates for the parameters b_0 and b_1 of the least-squares straight line in the output from the program.

(c) Write the model equation for the relationship between the amount of the drug received and the drop in blood pressure as it is written in part (b) of exercise 1 except use the actual numerical values of b_0 and b_1 in the equation (as you obtained in part (b) immediately above) instead of b_0 and b_1.

(d) Use the equation you wrote in part (c) to predict the drop in blood pressure for a patient in the population for whom the amount of the drug received is zero milligrams per day. Use the equation to predict the drop in blood pressure for a patient in the population for whom the amount of the drug received is 100 milligrams per day.

NOTE: The foregoing two predictions are not based on the assumption that a straight line is appropriate for the data because (since they use the same values of the predictor variable as were actually used in the experiment) these two predictions are based directly on the results of the experiment (as opposed to being based on "interpolation" or "extrapolation" of the results on the basis of the best-fitting line).

(e) Plot the two predicted values you obtained in part (d) on the scatterplot. [Use the plotting procedure in exercise 1 (e).] On the basis of the results of the experiment, these two points are the best predictions for the drop in blood pressure for patients in the population who receive (under circumstances similar to those of the experiment) the respective amount of the two amounts of the drug used in the experiment.

NOTE: It is, of course, to be expected that each of the two predictions lies in the middle of its associated cloud of points. In fact, the prediction algorithm works in such a way that each of these least-squares predictions is (regardless of what numerical values appear in table 4) simply equal to the average of the drops in blood pressure for the subset of patients in the research project who received the associated amount of the blood pressure drug. More generally, in actual experiments (but not in observational research projects) the mathematically "best" predictions (provided certain often-satisfiable assumptions are adequately satisfied) are simply group averages. Therefore, we often need not generate model equations in experiments in order to make predictions. Instead, (once we have verified that a relationship is present) we can simply predict using the appropriate group averages. A model equation was generated in this exercise to illustrate how the procedures for making standard predictions from observational data and the procedures for making standard predictions from experimental data are, although often implemented differently, effectively equivalent. These equivalent procedures are part of a broad set of methods for studying relationships between variables that are based on a "general linear model" (equation). The term "linear" is used to indicate that the methods are based on "straight line" relationships, although extensions to the methods allow them to effectively deal with a wide variety of nonlinear relationships.

(f) Find the statistic in the computer output labeled the "standard error (of the predictions)" or labeled with one of the other names listed in exercise 1 (i). As always, the standard error of the predictions is measured in the same units as the predictor variable. However, in the case of the experiment, twice the standard error is a good estimate of the size of a 68 percent confidence interval (centered on the predicted value) for individual predictions of the drop in blood pressure for new patients provided the conditions specified in exercise 1 (i) are satisfied, and provided that the new patients are administered one of the two amounts of the blood pressure drug that were used in the experiment. That is, if we administer one or the other of the two amounts of the drug to new patients from the population (and if the conditions are satisfied), then we can expect that roughly 68 percent of the predictions of the value of the drop in blood pressure for the new patients will be no more than the size of the standard error away from (above or below) the corresponding true measured values of the drop in blood pressure for the patients.

12. MISCELLANEOUS TECHNIQUES

The preceding three sections respectively describe statistical techniques for *detecting* relationships between variables, *illustrating* relationships between variables, and making *predictions* or exercising *control* on the basis of relationships between variables. A fourth group of statistical techniques for studying relationships between variables consists of various less-frequently-used but sometimes valuable techniques. For example:

- Techniques are available to perform (multiple) comparisons of

  the average(s) of the values
  of a response variable
  for one or more (sets of) values
  of the predictor variable(s)
  against
  the average(s) of the values
  of the response variable
  for one or more other (sets of) values
  of the predictor variable(s).

  These comparisons help us to determine whether the differences between the averages are greater than could be expected by chance, assuming that there is no relationship between the response variable and the predictor variables. These techniques help to illuminate particular aspects of the overall relationship between variables that we are studying.
- Techniques are available to calculate estimates of the "proportion of the variation" in the values of a response variable in a relationship that can be associated with "variation" in the values of one or more predictor variables (called "variance component analysis").
- Similarly, techniques are available to calculate measures of the "strength" of a relationship between a response variable and one or more predictor variables (e.g., a correlation coefficient).
- Techniques are available to test hypotheses about the values of the parameters in a model equation.
- Similarly, techniques are available to calculate confidence intervals for the estimates of the values of the parameters in a model equation.

13. THE ORDER OF USING THE TECHNIQUES

The preceding four sections discussed four groups of statistical techniques for studying relationships between variables. This raises the following question: If we are studying a relationship between variables, is there a preferred order of using the four groups of techniques?

The answer is that we should first use appropriate techniques to *detect* whether there is a relationship between the variables. Detecting should come first because the other techniques are mostly based on the assumption that there is a relationship. Thus usually it is ill-advised to attempt to use any of the other techniques on a set of data unless we have first detected clear evidence of a relationship between the variables in the data.

Second, if we have detected a relationship between variables, we should next use appropriate techniques to *illustrate* the relationship because illustrating a relationship will almost always help us to understand it.

Third, once we have detected and illustrated a relationship between variables, we are then often interested in deriving a method that will allow us to accurately predict or control the values of the response variable in new entities in the population on the basis of the values of the predictor variables in the entities.

14. THE ITERATIVE NATURE OF SCIENCE

After we have carried out one or more of the four groups of techniques for studying relationships between variables (and if we have found new information about a relationship between variables), we will have advanced our knowledge of relationships between variables in some area of experience. However, such advancement is almost never an end point. Instead, we usually wish to refine our knowledge of a relationship, often by involving more predictor variables in the relationship or by exploring the relationship when different values of the predictor variables are used. Scientific research is thus highly iterative (Box and Draper 1987, sec. 1.3).

15. SUMMARY AND NEXT STEPS

This paper has introduced the entity-property-relationship approach to science and statistics. We began by discussing the concepts of entities, properties of entities, and values of properties of entities. Then we discussed a main goal of science, which is to predict and control the values of properties of entities. The concept of a relationship between properties of entities was introduced next as a key to predicting and controlling the values of properties. Finally, four groups of statistical techniques were introduced to help study relationships between properties:

- techniques for *detecting* relationships between properties
- techniques for *illustrating* relationships between properties
- techniques for *predicting* and *controlling* the values of properties, and
- miscellaneous techniques for studying relationships between properties.

In order to provide an overview of the use of statistical techniques in scientific research, the discussion in this paper has sometimes been at an abstract level. If you wish to reinforce the concepts, it is recommended that you now study specific statistical methods that perform the four groups of techniques. For maximum practical understanding, it is recommended that you seek discussions of the methods that focus on solving realistic scientific problems.

APPENDIX A: GENERALIZATIONS OF THE CONCEPT OF *NULL HYPOTHESIS*

Scientists sometimes usefully extend the concept of a null hypothesis to include statements that claim there is no relationship in entities between a response variable and one or more predictor variables *beyond some already-accepted relationship.* If (as is often the case) the response variable under study is a continuous numeric variable, this type of null hypothesis can be easily translated into a standard hypothesis of "no relationship between variables" by "transforming" the values of the response variable. Under the transformation, the new value of the response variable for an entity is the *residual* obtained by subtracting the predicted (by the already-accepted model equation) value of the response variable for the entity from the observed value of the response variable for the entity. If we wish to reject a null hypothesis of this type, we must find reliable evidence of a relationship between the new transformed (residual) variable and the predictor variable(s).

We can usefully extend the concept of a null hypothesis still further to include all statements that claim that some entity or type of entity *does not exist.* Examples of such statements include statements of the non-existence of a particular relationship, or the non-existence of a difference between some population parameter and some constant, or the non-existence of nineteenth-century science's "luminiferous ether".

Of course, (provided, as usual, the "size" of the entity is not stated), even if a particular null hypothesis is true, it is impossible to prove conclusively that the hypothesis is true because it is impossible to prove conclusively that some entity that is logically possible does not exist. (It is interesting to note that the preceding sentence, being itself a null hypothesis, is also impossible to prove true.) On the other hand, if we look and succeed in reliably finding one or more instances of a particular type of entity, it clearly *is* possible to empirically prove, beyond a reasonable doubt, that that particular entity or type of entity (e.g., a particular relationship between variables) *does* exist. As discussed in subsection 9.2, the principle of parsimony im-

plies that we should begin the study of some phenomenon by assuming the simplest possible situation, which amounts to assuming various (unprovable) null hypotheses about the phenomenon. We can then perform research projects that attempt to refute these null hypotheses and thereby obtain knowledge about the phenomenon. This knowledge helps us to understand the external world.

APPENDIX B: FACTORS THAT AFFECT THE POWER OF STATISTICAL TESTS FOR DETECTING RELATIONSHIPS BETWEEN VARIABLES

Given our interest in detecting relationships between variables, it is useful to identify those factors that we can control in a research project in order to maximize the power of the statistical tests for detecting relationships. Some of these factors follow:

- *Predictor Variables Measured.* Power depends on how many of the variables related to the response variable are actually measured in a research project. (Often several relevant predictor variables are omitted from a research project because scientists are either unaware that these variables exist, or they are unaware that these variables are related to the response variable.) The more relevant predictor variables that we include in the analysis of a research project, the more powerful the statistical tests. (A predictor variable is "relevant" if it is at least partially independent of the other predictor variables in the research project and if the strength of the relationship between the response variable and this predictor variable is high enough that its inclusion in the analysis causes power to increase.) On the other hand, including *irrelevant* predictor variables in the analysis of the results of a research project usually causes the power of the statistical tests to slightly *decrease.*
- *Broadness of Variation in the Relevant Predictor Variables.* Power depends on the amount of variation in the values of the relevant predictor variables. Within the range of values of the predictor variables in which a relationship is "jointly monotonic" (which is often a relatively broad range), the broader the variation in the values of the relevant predictor variables in a research project, the more powerful the statistical tests.
- *Choice of the Number of Values for Each Relevant Manipulated Predictor Variable.* As noted at the end of subsection 7.9, if we manipulate the values of one or more of the predictor variables in a research project, the research project is called an *experiment.* In any experiment we must choose the number of values that we wish to cause each manipulated predictor variable to assume. Generally, for a fixed range of values in which the relationship is monotonic, the fewer (to a minimum of two) values that we cause a manipulated predictor variable to have in an experiment, the more powerful the statistical tests for relationships between this predictor variable and the response variable.
- *Sample Size.* Power depends on the number of entities in the sample in which we measure the values of the variables. Generally, the more entities in the sample, the more powerful the statistical tests.
- *Frequency of Measurement.* Power depends on the number of times that we measure the response variable and the relevant predictor variables in each entity in the sample. Generally, the more times we measure the set of variables in each entity (especially if the predictor variables are allowed or caused to vary broadly within each entity), the more powerful the statistical tests.
- *Value Allocation.* Power depends on how the different values of the relevant predictor variables are allocated to the different instances of measurement in the research project, as determined by the details of the design of the research project. Generally, the more "balanced" the allocation of the different (combinations of) values of the relevant predictor variable(s) to the different instances of measurement, the more powerful the statistical tests.
- *Similarity of Conditions.* Power depends on the similarity of the conditions under which the different entities in the sample participate in the research project. (Here, "similarity of conditions" means similarity across the research project of the values of variables that are *not* predictor variables in the research project.) Generally, the more similar the conditions throughout the research project, the more powerful the statistical tests. (Power usually increases with similarity of conditions because similar conditions imply that the values of relevant unmeasured predictor variables are more likely to be similar across the research project, thereby reducing the size of the "error term" in the statistical tests, and thereby increasing the power.)
- *Homogeneity of Sample.* Power depends on how similar the entities in the sample are to each other. Generally, the more similar the entities (i.e., the more homogeneous the sample), the more powerful the statistical tests. (Power usually increases with homogeneity because highly similar entities are usually also similar in the values of relevant unmeasured predictor variables, thereby reducing the size of the error term.) Choosing highly similar entities for participation in a research project is a reasonable research strategy for increasing power, although it does have one drawback: If we find a relationship then, strictly speaking, we can only generalize the relationship to the population of (highly similar) entities from which we selected the sample, rather than to a more general population of entities. However, this loss of generality is sometimes acceptable in order to increase the likelihood of finding a relationship. Once we have found a useful relationship, we can then, if neces-

sary, do further research to examine its generality.

- *Choice of Statistical Test.* In some research situations more than one statistical test can be validly used to look for evidence of a particular type of relationship between the response variable and a set of predictor variables. (Formally, to minimize the false alarm error rate, in any particular research project and for any particular type of relationship being sought in that research project—e.g., a two-way interaction—we should usually use only one of the available tests.) Generally, the different statistical tests will have different powers, and generally one of the available tests can be shown to have the highest possible power for detecting the expected form of the relationship.
- *Choice of Measurement Type.* Methods for measuring the values of variables in entities can be classified as being either "continuous" or "discrete". (Continuous measurement methods can theoretically return an unlimited number of different values, whereas discrete methods can return only a limited number of different values, often less than ten.) Generally, in situations in which we can measure the values of the response variable or a relevant predictor variable with both a continuous measurement method and a discrete measurement method, the most powerful available statistical test using the continuous method will be more powerful than the most powerful available test using the discrete method.
- *Measurement Accuracy.* Regardless of whether we use continuous or discrete measurement methods, power depends on the accuracy (reliability) of the methods we use to measure the values of the response variable and the relevant predictor variables. Generally, the more accurate the measurement methods, the more powerful the statistical tests.
- *Choice of Critical* p-*Value.* Power depends on the value we choose for the critical *p*-value: the higher the critical *p*-value, the more powerful the statistical tests. Of course, we are generally not free to set the critical *p*-value to an arbitrarily high value (e.g., .5) because this causes a high rate of false alarms. Instead, as noted in subsection 9.4, many scientists find it reasonable to consistently use a critical *p*-value of .05 or .01.

Further discussion of these and other factors that affect the power of statistical tests is given in many of the works listed at the end of subsection 11.1.

APPENDIX C: MULTI-HYPOTHESIS RESEARCH PROJECTS

Rather than having a single clear hypothesis (or a very few clear hypotheses) to test, some research projects are designed to test for the existence of many possible relationships between one or more response variables of interest and perhaps hundreds of predictor variables. For example, in medical screening for the effects of new drug compounds, hundreds of different compounds may be tested in different parallel experiments to see if any of these compounds have an effect on several important response variables. We shall call these research projects "multi-hypothesis" research projects because they use many statistical tests to test many different hypotheses.

Now, if we use a critical *p*-value of, say, .05 in a multi-hypothesis research project, then in five percent of the statistical tests in which there is no relationship we will erroneously conclude that the relationship exists. Therefore, because so many hypotheses are being tested, we are almost sure to commit some false alarm errors. Therefore, most statisticians and researchers feel that in this situation the usual decision rule is too lax. Therefore, statisticians have defined procedures to tighten the criteria for the conclusion that a relationship exists in multi-hypothesis research projects. The procedures are discussed under the headings of multiple comparisons, simultaneous inference methods, multiple testing, and multiplicity. Hochberg and Tamhane (1987), Miller (1981), and Tukey (1993) discuss these procedures, which amount to various ways of effectively lowering the critical *p*-value for accepting the existence of a relationship.

Although these procedures are sometimes useful, a problem with them is that they severely diminish the power of the statistical tests for detecting relationships. Thus some scientists adopt another approach, which does not diminish the power of the tests and which is often relatively easy to implement in a multi-hypothesis research project. The approach is simply to independently repeat any portion of the research project in which a *p*-value of less than, say, .05 is obtained for a relationship. And (provided that the number of repeated research projects is not excessive) if a *p*-value for the relationship of less than .05 is validly obtained in the repeated version of a research project, it is then quite reasonable to conclude that the associated relationship between variables probably exists in the population.

APPENDIX D: CHOOSING THE VALUES OF THE PARAMETERS IN THE MODEL EQUATION WHEN THE RESPONSE VARIABLE IS CONTINUOUS

The response variable in a research project is often continuous—that is, it can theoretically have any value within some continuous range of numerical values. When the response variable in a research project is continuous, most approaches that are designed to minimize errors in predictions work by choosing the values of parameters in the model equation so as to mathematically minimize the sum (across all the residuals—defined in subsection 11.1) of the values of some monotonically increasing function of the absolute size of the residuals. One popular function is the absolute value function itself, as proposed by Bo-

scovich in the middle of the eighteenth century (Eisenhart, 1961). Another popular function is the squaring function ("least squares"), as proposed by Gauss and Legendre at the beginning of the nineteenth century (Plackett, 1972). The absolute value function has the advantage of appropriately assigning "outliers" less relative weight than the squaring function when determining the estimates of the values of the parameters. However, the squaring function has the advantages of being mathematically easier to deal with, and (under often-satisfiable assumptions) it minimizes the "variance" of the predictions made by the model equation. (However, despite the high level of acceptance of variance as a measure of spread [with the acceptance probably mainly due to the high mathematical tractability of variance and related concepts], the use of variance as a measure of spread is, although reasonable, arbitrary.) Once the form of the model equation is chosen, and once the appropriate function of the residuals is chosen, it is a simple exercise in elementary calculus or computer iteration to derive the estimates of the values of the parameters. Draper and Smith (1981, sec. 1.2) describe the calculus for a commonly-used approach.

APPENDIX E: FACTORS THAT AFFECT PREDICTION AND CONTROL ACCURACY

For economy of expression, this appendix uses the concept of *prediction* to stand for the concept of *prediction or control.*

When we use statistical prediction methods we usually want the predictions to be as accurate as possible. Therefore, it is useful to identify the factors we can control in a research project to maximize the accuracy of any subsequent predictions we attempt on the basis of a relationship between variables discovered in the research project. Some of the factors that affect prediction accuracy follow:

- *Representativeness of Sample.* Generally, the more representative the sample of entities in the original research project is of the population for whom we wish to make predictions, the more accurate the predictions. We can maximize the representativeness of a sample by using a random sampling technique to select the sample from the population.
- *Predictor Variables Measured.* Generally, the more relevant predictor variables we measure in the original research project (assuming, of course, that we measure the same predictor variables with the same measurement methods in the prediction situation), the more accurate the predictions.
- *Sample Size.* Generally, the more entities in the sample in the original research project, the more accurate the predictions.
- *Frequency of Measurement.* Generally, the more times we measure the values of the response variable and the relevant predictor variables in each entity in the original research project, the more accurate the predictions.
- *Manipulation.* When designing a research project we can choose either (*a*) to simply observe the values of the predictor variables in the entities in the research project or we can choose (*b*) to actively manipulate the values of some, or all, of the predictor variables in the entities. If we wish to learn how to *control* the values of the response variable in new entities (as opposed to merely learning how to *predict* the values), it is useful to manipulate the predictor variables because control capability cannot be assured without manipulation. Manipulation of predictor variables in experiments is discussed further in books about experimental methods and by Macnaughton (1996b, chap. 4).
- *Value Allocation.* Generally, the more balanced the allocation of the different (combinations of) values of the relevant predictor variable(s) to the different instances of measurement, the more accurate the predictions will be overall.
- *Similarity Between Research Situation and Prediction Situation.* Prediction accuracy depends on the similarity between (*a*) the values of the relevant predictor variables (including both measured and unmeasured variables) in the original research project and (*b*) the values of the same predictor variables in the situation in which the predictions will be made. Generally, the more similar the values of the relevant predictor variables in the two situations, the more accurate the predictions. We can maximize this similarity by making all the physical conditions in the original research project as similar as possible to the conditions under which the predictions will later be made. In an experiment (as opposed to an observational research project), we can help to maximize the similarity between the experiment and the prediction situation by using more than two values for each manipulated predictor variable in the experiment. In such cases, unless it is known that most predictions will be made in a certain part of the range, it is often best to spread the different values evenly throughout the chosen range. (Although using more than two values for a relevant manipulated predictor variable in an experiment may increase prediction accuracy, using more than two values will also usually *decrease* the power of the associated statistical tests, as noted in appendix B. Thus in designing an experiment the researcher must decide which is more important: prediction accuracy or power of statistical tests. The necessity of this compromise sometimes leads researchers to use three values for each manipulated predictor variable: two extreme values and a "center point".)
- *Confounding.* Prediction accuracy can depend on the amount of "confounding" between the relevant predictor variables (both measured and unmeasured) in a research

project. (Two predictor variables are *confounded* in a research project if the values of the two variables are allowed or caused to vary fully or partially "in step" with each other in the entities in the research project.) If two predictor variables are confounded, and if a relationship is found between the response variable and these predictor variables, we may be unable to tell which of the two predictor variables is more directly related to the response variable. Thus if two predictor variables are confounded in a research project, we may be unable to tell which one is the better predictor of the response variable (or whether some combination of the two variables is best). In an observational research project, once we have chosen the sample and the predictor variables we usually do not have direct control over confoundings between the predictor variables—we must take what we get. On the other hand, in an experiment, we can minimize the possibility of confoundings between predictor variables by (*a*) designing the experiment so that there are no confoundings between any pairs (or larger groups) of the manipulated (predictor) variables (i.e., by ensuring that all the manipulated variables are caused to vary independently of each other in the entities) and by (*b*) randomly assigning the entities in the experiment to the different treatment groups, thereby minimizing the chance that confoundings will occur between the manipulated variables and other relevant variables. Fisher identified the fundamental importance of random assignment of entities to treatment groups in experiments (Box 1978, pp. 144-152).

- *Choice of Measurement Type.* Generally, in situations in which we can measure the values of a response or predictor variable with both a continuous measurement method and a discrete measurement method, the most accurate available predictions using the continuous method will be more accurate than the most accurate available predictions using the discrete method.
- *Measurement Accuracy.* Generally, the more accurate (reliable) the measurement methods for the response and relevant predictor variables in the original research project (assuming that we measure the same variables with the same measurement methods in the prediction situation), the more accurate the predictions.
- *Choice of Prediction Method.* In some research situations more than one prediction method can be validly used to predict the value of a response variable on the basis of the values of a set of predictor variables. Generally, the different prediction methods will yield different prediction accuracies, and generally, for a particular prediction situation, one of the available methods can be shown to be the most accurate.

Further discussion of these and other factors that affect prediction accuracy is given in many of the works listed at the end of subsection 11.1.

REFERENCES

Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley.

Anderson, T. W. (1971), *The Statistical Analysis of Time Series*, New York: John Wiley. (Republished by John Wiley in 1994.)

Anderson, R. L., and Bancroft, T. A. (1952), *Statistical Theory in Research*, New York: McGraw-Hill.

Anscombe, F. J., and Tukey, J. W. (1963), "The Examination and Analysis of Residuals," *Technometrics*, 5, 141-160.

Barnett, V. (1988), "Relationship," in *Encyclopedia of Statistical Sciences* (Vol. 8), ed. S. Kotz and N. L. Johnson, New York: John Wiley, pp. 12-14.

Barnett, V., and Lewis, T. (1994), *Outliers in Statistical Data* (3rd ed.), Chichester, England: John Wiley.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis* (2nd ed.), New York: Springer-Verlag.

Bernardo, J. M., and Smith, A. F. M. (1994), *Bayesian Theory*, Chichester, England: John Wiley.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.

Bollen, K. A. (1989), *Structural Equations with Latent Variables*, New York: John Wiley.

Box, G. E. P., and Draper, N. R. (1987), *Empirical Model Building and Response Surfaces*, New York: John Wiley.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978), *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, New York: John Wiley.

Box, G. E. P., and Jenkins, G. M. (1970), *Time Series Analysis: Forecasting and Control*, San Francisco, CA: Holden-Day.

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley. (Republished in 1994 by John Wiley in New York.)

Box, J. F. (1978), *R. A. Fisher, The Life of a Scientist*, New York: John Wiley.

Chatterjee, S., and Price, B. (1991), *Regression Analysis by Example* (2nd ed.), New York: John Wiley.

Chowdhury, A. R. (1987), "Are Causal Relationships Sensitive to Causality Tests?" *Applied Economics*, 19, 459-465.

Cleveland, W. S. (1993), *Visualizing Data*, Summit, NJ: Hobart Press.

Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: John Wiley.

Cochran, W. G., and Cox, G. M. (1957), *Experimental Designs* (2nd ed.), New York: John Wiley. (Republished by John Wiley in 1992.)

Cox, D. R. (1958), *Planning of Experiments*, New York:

John Wiley. (Republished by John Wiley in 1992.)

Cox, D. R., and Snell, E. J. (1989), *Analysis of Binary Data* (2nd ed.), London: Chapman and Hall.

Cyert, R. M., and DeGroot, M. H. (1987), *Bayesian Analysis And Uncertainty In Economic Theory*, Totawa, NJ: Roman and Littlefield.

Dawid, A. P. (1986), "A Bayesian View of Statistical Modelling," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, ed. P. K. Goel and A. Zellner, Amsterdam: North-Holland, pp. 391- 404.

Deming, W. E. (1950), *Some Theory of Sampling*, New York: John Wiley.

Draper, N. R., and Smith, H. (1981), *Applied Regression Analysis* (2nd ed.), New York: John Wiley.

Eisenhart, C. (1961), "Boscovich and the Combination of Observations," in *Roger Joseph Boscovich*, ed. L. L. Law, London: Allen & Unwin.

Fienberg, S. E. (1980), *The Analysis of Cross-Classified Categorical Data* (2nd ed.), Cambridge, MA: MIT Press.

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd. The 14th edition of this seminal work appears in Fisher (1990).

—— (1935), *The Design of Experiments*, Edinburgh: Oliver and Boyd. The 8th edition of this seminal work appears in Fisher (1990).

—— (1990), *Statistical Methods, Experimental Design, and Scientific Inference*, ed. J. H. Bennett, Oxford: Oxford University Press.

Freedman, D., Pisani, R., Purves, R., and Adhikari, A. (1991), *Statistics* (2nd ed.), New York: Norton.

Good, I. J. (1992), "The Bayes/Non-Bayes Compromise: A Brief Review," *Journal of the American Statistical Association*, 87, 597- 606.

Goodman, L. A. (1984), *The Analysis of Cross-Classified Data Having Ordered Categories*, Cambridge, MA: Harvard University Press.

Goodman, L. A., and Kruskal, W. H. (1979), *Measures of Association for Cross Classifications*, New York: Springer-Verlag.

Granger, C. W. J. (1980), "Testing for Causality: A Personal Viewpoint," *Journal of Economic Dynamics and Control*, 2, 329-352.

Graybill, F. A. (1976), *Theory and Application of the Linear Model*, North Scituate, MA: Duxbury.

Haberman, S. J. (1978-1979), *Analysis of Qualitative Data* (2 vols.), New York: Academic Press.

Hahn, G. J., and Meeker, W. G. (1991), *Statistical Intervals: A Guide for Practitioners*, New York: John Wiley.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953), *Sample Survey Methods and Theory* (2 vols.), New York: John Wiley. (Republished by John Wiley in 1993.)

Hartigan, J. A. (1983), *Bayes Theory*, New York: Springer-Verlag.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1983), *Understanding Robust and Exploratory Data Analysis*, New York: John Wiley.

—— (1985), *Exploring Data Tables, Trends, and Shapes*, New York: John Wiley.

—— (1991), *Fundamentals of Exploratory Analysis of Variance*, New York: John Wiley.

Hochberg, Y. and Tamhane, A. C. (1987), *Multiple Comparison Procedures*, New York: John Wiley.

Hocking, R. R. (1985), *The Analysis of Linear Models*, Monterey, CA: Brooks-Cole.

Howson, C., and Urbach, P. (1993), *Scientific Reasoning: The Bayesian Approach* (2nd ed.), Chicago: Open Court.

Huizenga, J. R. (1993), *Cold Fusion: The Scientific Fiasco of the Century*, New York: Oxford University Press.

Iman, R. L. (1994), *A Data-Based Approach to Statistics*, Belmont, CA: Wadsworth, Duxbury Press.

Jöreskog, K. G. (1993), "Testing structural equation models," in *Testing Structural Equation Models*, ed. K. A. Bollen and J. S. Long, Newbury Park, CA: Sage, pp. 294-316.

Kempthorne, O. (1952), *The Design and Analysis of Experiments*, New York: John Wiley.

Kendall, M., and Ord, J. K. (1990), *Time Series* (3rd ed.), London: Hodder and Stoughton.

Kendall, M., Stuart, A., and Ord, J. K. (1987), *Kendall's Advanced Theory of Statistics* (Vol. 1), London: Griffin.

Kish, L. (1965), *Survey Sampling*, New York: John Wiley.

Kleinbaum, D. G., Kupper, L. L., and Muller, K. E. (1988), *Applied Regression Analysis and Other Multivariate Methods* (2nd ed.), Boston: PWS-Kent.

Koch, G. G., Carr, G. J., Amara, I. A., Stokes, M. E., and Uryniak, T. J. (1990), "Categorical Data Analysis," in *Statistical Methodology in the Pharmaceutical Sciences*, ed. D. A. Berry, New York: Marcel Dekker, pp. 389-473.

Kraemer, H. C. and Thiemann, S. (1987), *How Many Subjects?* Newbury Park, CA: Sage.

Lee, P. M. (1989), *Bayesian Statistics: An Introduction*, New York and Oxford: John Wiley; London: Edward Arnold.

Lehmann, E. L. (1986), *Testing Statistical Hypotheses* (2nd ed.), New York: John Wiley.

Lindley, D. V. (1983), "Theory and Practice of Bayesian Statistics," *The Statistician*, 32, 1-11.

Lipsey, M. W. (1990a), "Theory as Method: Small Theories of Treatments," in *Research Methodology: Strengthening Causal Interpretations of Nonexperimen-*

tal Data, ed. L. Sechrest, E. Perrin, and J. Bunker, Rockville MD: U.S. Department of Health and Human Services, pp. 33-51.

Lipsey, M. W. (1990b), *Design Sensitivity*, Newbury Park, CA: Sage.

Macnaughton, D. B. (1996a), "The Introductory Statistics Course: A New Approach." To be submitted for publication. This 8000-word draft paper is available at *http://www.matstat.com/teach*

Macnaughton, D. B. (1996b), *How to Predict the Unknown: An Introduction to Scientific Thinking With Statistics*, textbook, in preparation.

Mason, R. L., Gunst, R. F., and Hess, J. L. (1989), *Statistical Design and Analysis of Experiments*, New York: John Wiley.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Miller, R. G. Jr. (1981), *Simultaneous Statistical Inference* (2nd ed.), New York: Springer-Verlag.

Montgomery, D. C., and Peck, E. A. (1992), *Introduction to Linear Regression Analysis* (2nd ed.), New York: John Wiley.

Moore, D. S. (1995), *The Basic Practice of Statistics*, New York: W. H. Freeman.

Mosteller, F. (1968), "Association and Estimation in Contingency Tables," *Journal of the American Statistical Association*, 63, 1-28.

—— (1990), "Improving Research Methodology: An Overview," in *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data*, ed. L. Sechrest, E. Perrin, and J. Bunker, Rockville MD: U. S. Department of Health and Human Services, pp. 221-230.

Mosteller, F., and Tukey, J. W. (1977), *Data Analysis and Regression*, Reading MA: Addison-Wesley.

Mosteller, F., and Wallace, D. L. (1984), *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*, New York: Springer-Verlag.

Myers, R. H., and Montgomery, D. C. (1995), *Response Surface Methodology*, New York: John Wiley.

Nelder, J. A., and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society A*, 135, 370-384.

Neyman, J., and Pearson, E. S. (1928), "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference, Part I," *Biometrika*, 20A, 175-240.

Odeh, R. E., and Fox, M. (1991), *Sample Size Choice* (2nd ed.), New York: Marcel Dekker.

Plackett, R. L. (1972), "Studies in the History of Probability and Statistics. XXIX The discovery of the method of least squares," *Biometrika*, 59, 239-251.

—— (1981), *The Analysis of Categorical Data*, London: Griffin.

Poirier, D. J. (1988), "Causal Relationships and Replicability," *Journal of Econometrics*, 39, 213-234.

Press, J. S. (1989), *Bayesian Statistics: Principles, Models, and Applications*, New York: John Wiley.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications* (2nd ed.), New York: John Wiley.

Scheffé, H. (1959), *The Analysis of Variance*, New York: John Wiley.

Searle, S. R. (1987), *Linear Models for Unbalanced Data*, New York: John Wiley.

Snedecor, G. W., and Cochran, W. G. (1989), *Statistical Methods* (8th ed.), Ames, IA: Iowa State University Press.

Tufte, E. R. (1983), *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press.

—— (1990), *Envisioning Information*, Cheshire, CT: Graphic Press.

Tukey, J. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

—— (1984), *The Collected Works of John W. Tukey: Volume I: Time Series 1949-1964*, ed. D. R. Brillinger, London: Chapman and Hall.

—— (1985), *The Collected Works of John W. Tukey: Volume II: Time Series 1965-1984*, ed. D. R. Brillinger, London: Chapman and Hall.

—— (1988), *The Collected Works of John W. Tukey: Volume V: Graphics 1965-1985*, ed. W. S. Cleveland, London: Chapman and Hall.

—— (1989), "SPES in the Years Ahead," in *Proceedings of the American Statistical Association Sesquicentennial Invited Paper Sessions*, Alexandria, VA: The American Statistical Association, pp. 175-182.

—— (1992), *The Collected Works of John W. Tukey: Volume VII: Factorial and ANOVA, 1949-1962*, ed. D. R. Cox, London: Chapman and Hall.

—— (1993), *The Collected Works of John W. Tukey: Volume VIII: Multiple Comparisons*, ed. H. I. Braun, London: Chapman and Hall.

Weisberg, S. (1985), *Applied Linear Regression* (2nd ed.), New York: John Wiley.

Winer, B. J. (1971), *Statistical Principles in Experimental Design*, New York: McGraw-Hill.

Winkler, R. L. (1972), *An Introduction to Bayesian Inference and Decision*, New York: Holt, Rinehart & Winston.

Yates, F. (1981), *Sampling Methods for Censuses and Surveys* (3rd ed.), London: Griffin.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, New York: John Wiley. (Republished in 1987 by Krieger in Melbourne, FL.)

(version of April 25, 1996; revised on November 11, 1997 to correct Internet addresses and typos only)

Mathematical expressions in this document were generated with MathType by Design Science, *http://www.mathtype.com/mathtype/*

CONTENTS