THEORIES OF DATA ANALYSIS
FROM MAGICAL THINKING THROUGH CLASSICAL STATISTICS

by
Persi Diaconis
Stanford University

TECHNICAL REPORT NO. 206
OCTOBER 1983

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT
MCS80-24649

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

THEORIES OF DATA ANALYSIS
FROM MAGICAL THINKING THROUGH CLASSICAL STATISTICS

by
Persi Diaconis
Stanford University

TECHNICAL REPORT NO. 206
OCTOBER 1983

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT
MCS80-24649

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Theories of Data Analysis: From Magical Thinking through Classical Statistics

1A. Intuitive Statistics -- Some Inferential Problems

1B. Multiplicity -- A Pervasive Problem

1C. Some Remedies

1D. Theories for Data Analysis

1E. Uses for Mathematics

1F. In Defense of Controlled Magical Thinking

Exploratory data analysis (EDA) is the art of finding structure, or simple descriptions, in data. We look at numbers or graphs and try to find patterns. We are encouraged to pursue leads suggested by background information, imagination, the patterns perceived, and our experience with other data analyses.

Magical Thinking

Magical thinking involves our inclination to seek and interpret connections between the events around us together with our disinclination to revise belief after further observation. In one manifestation, it may be believed that the Gods are sending signs, and that a particular ritual holds the key to understanding. This belief persists despite the facts. Of course, this is but one aspect of a rich collection of facts and theories describing man's myths and rituals. Chapter 7 of Cassirer (1972) is an inspring presentation of the role of magical thinking in modern life, Schweder (1977) describes other anthropological and psychological studies.

The south sea Cargo cults are a clear example of magical thinking. During World War II, airplanes landed and unloaded food and materials. To bring this about again, the natives build fires along the sides of makeshift runways. They have a ceremonial controller who sits in a hut with a wooden helmet sporting bamboo bar antennas. They recreate the pattern observed in the past and wait for the planes to land. No airplanes land, yet the ritual continues. Worsley (1960) offers documentation and an anthropological interpretation.

Several studies of human learning indicate that such behavior is common in our own activities, In a typical study, an experimenter generates a pattern of zeros and ones by flipping a coin (with chances of heads .7) Subjects are asked

to predict successive outcomes, not knowing the chance of generating mechanism. The subjects seek and see patterns (e.g., alternation or a head following two tails) and often report a complex working system even after hundreds of trials. Subjects do not learn and converge to the right strategy (always guess heads).

At one extreme, we can view the techniques of EDA as a ritual designed to reveal patterns in a data set. Thus, we may believe that naturally occurring data sets contain structure, that EDA is a useful vehical for revealing the structure, and that revealed structure can be interpretted in the language of the substantive subject matter that gave rise to the data. If we make no attempt to check if the structure could have arisen by chance; and tend to accept the findings as gospel, then the ritual comes close to magical thinking.

None of this argues that exploration is useless or wrong. Consulting statisticians cannot be universal subject matter experts and data often present striking patterns. Such explorations have been, and continue to be, an important part of science. In the final section of this chapter, it is argued that a controlled form of magical thinking--in the guise of "working hypothesis"--is a basic ingredient of scientific progress.

Classical Mathematical Statistics

Classical mathematical statistics offers a different description of what we do (or should do) when we examine data. It seeks to interpret patterns as chance fluctuations. In its most rigid formulation, we decide upon models and hypotheses before seeing the data. Then, we compute estimates and carry out tests of our assumptions. Classical statistics offers an antidote for some of the problems of magical thinking. However, as a description of what a real scientist does when confronting a real, rich data base, it seems as far off as a primitive ritual.

In some cases the standard errors and p-values of classical statistics are useful, even though the interpretation of "wrong once in 20 times under chance conditions" is not valid or interesting. Long experience has given practitioners a common understanding of the use and usefulness of classical procedures. Some of the surprising things a young statistician learns when working with an experienced applied statistician come in the form of statements like "oh, the overall F-test is always significant in problems like this, that's nothing to get excited about. It's the relative size of the t-tests for the contrasts that you have to watch." In these fields, standard errors and p-values have evolved as a useful way of communicating. Such uses can be studied; they are most worthy of respect.

Many areas of science and technology--the latter including medicine and agriculture, for example--are cooperative joint ventures of many people and many groups. One of the most important functions that formalized schemes of data analysis can play is to facilitate and deepen communication among these groups and individuals. Once the statistician gives up the hubris of being "the decider," he or she can seize a vital role, helping to make science and technology function better, because groups and people now understand the quantitative aspects of the strengths of each other's results more clearly. The qualitative aspects, frequently even more important, will have to be judged by those who know the field.

In one way classical statistics facilitates magical thinking: sometimes, the "person with the data" looks to "the statistician" as someone who will give the answers and take away the uncertainties, just as the primitive tribesman looks to the local shaman as one who will practice the rituals and take away the sickness. Tukey (1969) discusses the use of classical statistics as a ritual for sanctification.

Scientific Thinking

People have varied views of science. Is it a magnificent structure, erected to last forever, made up of irrefutable results? Some modern philosophy and history of science argues that this is not the case (e.g., Thomas Kuhn (1970) on scientific revolutions or Imre Lakatos (1976) on even so solid a discipline as mathematics). At another extreme, some creationists appear to believe that science is a thing made up of tissue and string. Surely this is too extreme--science has taken huge steps forward in the past few hundred years and, even though it errs and we cannot neatly say just what it is, there is clearly something there. John Tukey has said that "All the laws of physics are wrong, at least in some ultimate detail, though many are awfully good approximations." Some such middle ground must be correct.

In much of science, repetition on new data is both possible and practical. Such science proceeds by combining exploration with -- attempted confirmation on new data. Repetition with distinct data and distinct experimenters or observers is a hallmark that distinguishes science from magical thinking, although both share a substantial dependence on exploration.

Much of science also falls under John Tukey's label "uncomfortable science," because real repetition is not feasible or practical. Geology and geophysics are often on nonreplicable observation--fossil plate boundaries seem inadequately preserved to give us good replication of today's plate boundaries. Macroeconomics cannot find a second example of the Great Depression. Astronomical observation is often non-replicable.

Scientific thinking for uncomfortable science is not easy or simple to describe. It depends heavily, when practiced in the best style, on borrowing concepts, insights, and quantities from situations judged to be parallel, at least in specific aspects, to the situation at hand. Some examples are in remedy 4 of section 1c. As a result, changes in theory, which can involve new concepts or insights as well as new mathematical models, can have great impact on the conclusions drawn from a fixed set of observations. A frankly exploratory attiude seems mandatory when working with such data.

This chapter examines some theories of data analysis that fall between the extremes. The chapter reviews the

empirical work on intuitive statistics, suggests practical remedies for the most common problems, and goes on to review some of the frameworks erected especially for EDA. Uses for mathematics in EDA are discussed, and at the end a summary emphasizes the necessity of exploratory analysis and argues for the usefulness of a controlled form of magical thinking.

1A. INTUITIVE STATISTICS--SOME INFERENTIAL PROBLEMS

The skills, background information, and biases of a data analyst may affect the conclusions drawn from a body of data. People can imagine patterns in data. Sampling fluctuations can produce an appearance of structure. Without a standardized ritual, different investigators may come to different conclusions from the same evidence. This section surveys some studies that quantify these claims.

Studies of human perceptual abilities and human judgment shed some light on the subjective element in an exploratory analysis. Let us begin with a study that investigates how choices in drawing a scatter plot affect perception of the association between the two variables being plotted. Cleveland, Diaconis, and McGill (1982) compared two styles of plotting the same data. The first draws the scatter plot so that the points occupy a small part of the total space available. In the second version the points essentially fill the space available.

An example appears in Figure 1-1. Professional statisticians and other scientists with statistical training were asked to judge "how associated the two variables are." Most of the subjects judged a small plot as more associated than a big plot of the same points. The small plots are simply rescaled versions of the big plots; any of the standard measures of association (e.g., the correlation coefficient) is unchanged by rescaling. The subjective effects of rescaling are most

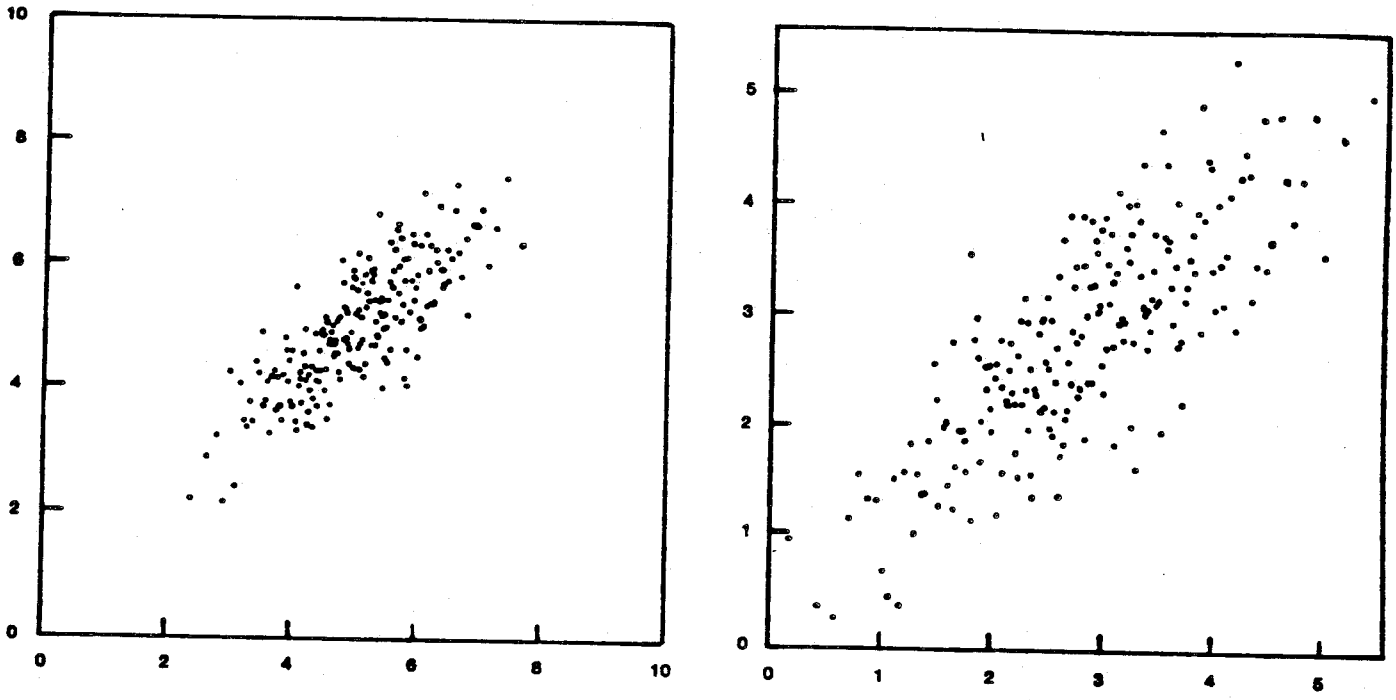
pronounced for data sets with correlation between .3 and .8. Rescaling can shift the perceived association by 10 to 15%. As studies like this become more popular, we can look forward to a better understanding of our psychological reaction to commonly used graphical procedures and comparisons between competing procedures.

Figure 1-1 about here

A different branch of psychology investigates how people make judgments under uncertainty. The book by Kahneman, Slovic, and Tversky (1982) assembles the basic experimental papers with extensive discussion. The book by Nisbett and Ross (1981) presents the basic research in a unified fashion, describing many replications and follow-up studies. Here we briefly describe some experiments that illustrate the degree to which experimenter preconception affects inference.

The anchoring phenomenon, described in Chapter 1 of Kahneman, Slovic, and Tversky (1982), exemplifies the effect of an experimenter's preliminary evaluations. Subjects were asked to estimate various percentages (e.g., the percentage of African countries in the United Nations). For each quantity, a number between 0 and 100 was determined by spinning a wheel in view of the subject. The subjects were instructed first to indicate whether that number was higher or lower than the actual value of the quantity and then to estimate that actual value by moving up or down from the given number. The starting number had a marked effect on estimates. For example, the median estimates

Figure 1-1. Scatterplots illustrating two styles of plotting data. Both scatterplots have $r = .8$.



Source: William S. Cleveland, Persi Diaconis, and Robert McGill (1982). "Variables on scatterplots look more highly correlated when the scales are increased," Science, 216, 1138-1141 (Figure 1 on page 1139).

of the percentage of African countries in the United Nations were 25 and 45 for groups of subjects that received 10 and 65, respectively, as starting points.

The anchoring phenomenon shows that a seemingly irrelevant random start can markedly affect the outcome. When people have a theory before encountering evidence, the theory takes precedence. Studies discussed in Chapter 8 of Nisbett and Ross suggest that data (whether they support the theory, oppose the theory, or are mixed) will tend to result in more belief in the original theory than seems normatively warranted.

In one study, Ross and Lepper (1980) presented Stanford University students with two purportedly authentic studies on the deterrent effects of capital punishment. The students had previously indicated that they either strongly believed capital punishment to be a deterrent to potential murderers or strongly believed it to be worthless as a deterrent. In a balanced design, each student read about the results and methods of an empirical study that demonstrated the effectiveness of capital punishment and about the results and methods of a study that demonstrated the lack of effect of capital punishment (the order was randomized).

Subjects found whichever study supported their own position to be significantly "more convincing" and "better conducted" than the study opposing their position. Subjects were asked about their beliefs after reading about only one study. Belief in initial position was strengthened if the study supported the subject's initial position. But belief in initial position

was affected relatively little if the study opposed the subject's initial position--subjects found flaws in the study design and conclusions in this circumstance. After reading about both studies--one that supported their initial position and one that opposed their initial position--the subjects were more convinced of the correctness of their initial position than they were before reading about any evidence. The evidence polarized students with differing views instead of bringing them closer together.

Another circle of studies quantifies the effect of experimenter bias. There are now many experiments showing that when experimenters expect certain responses from research subjects they are substantially more likely to observe what they expect. A summary of this work is in Rosenthal (1981).

The literature we have been discussing documents many other pitfalls that can befall intuitive statisticians. Even experienced analysts often believe in the "law of small numbers"; they think that a sample randomly drawn from a population is highly representative--similar to the population in all its essential characteristics (Chapter 2 of Kahneman, Slovic, and Tversky (1982)). People seem not to pay much attention to sample size, and the difference between a random sample and one selected to prove a point seems unimportant (Chapter 4 of Nisbett and Ross (1980)).

Availability and Representativeness

The psychologists who have undertaken these studies have done much more than point to normative aberrations. Often they

can explain the behavior of subjects on the basis of a few heuristics that we all seem to use. The two main heuristics, availability and representativeness, each deserve an example.

When using the availability heuristic, we judge population frequencies by the ease with which we can think of an example. For instance, Kahneman and Tversky showed that subjects asked to read lists of well-known personalities of both sexes subsequently overestimated the proportion of the sex on the list with the most famous names.

When using the representativeness heuristic, we attribute great weight to the resemblance or representativeness between the sample in hand and the population. For example, if subjects are asked to assess the relative likelihood of the following three birth order sequences of boys (B) and girls (G)

(1) BBBBBB (2) GBBGGB (3) GGGBBB

they think that (2) is most likely, followed by (3), followed by (1). If anything, (1) is actually most likely because of the slight preponderance of male births, and (2) and (3) are equally likely. But (2) seems representative of a random sequence.

Magical Thinking Again

One key difference between thinking and magical thinking is change in belief brought about by experience. If a theory or hypothesis does not hold up under replication, it should fall by the wayside. The psychological evidence just presented also shows that learning from experience can be very difficult. Subjects

with years of statistical training (including the present author) often err in much the same way as beginning students. This is particularly true if the problem comes up in a real life situation. The human learning studies cited in the introduction, and the Ross-Leper study on capital punishment make the same point: Once we notice a pattern or take hold of an idea, it can be harder than seems normatively warrented to revise our belief after further observation.

Data Analytic Examples

The studies described above are not so directly related to the daily work of the data analyst. Although suggestive, they are often conducted in carefully controlled circumstances on small, specific tasks. It has been argued that no serious scientist will be misled, in a large way, in a real study. Not with all the checks and balances of normal scientific protocol! Unfortunately, examples of faulty conclusions drawn from snooping about in data are all too easy to find. To illustrate, we mention three major examples.

Uncontrolled observations in medical trials. A doctor, or medical team, notices a pattern in available patient records and starts making medical recommendations on that basis. Studies surveyed in Bunker, Barnes, and Mosteller (1977) show the importance of--and need for--the elaborate machinery of randomized clinical trials and tests. Without it, optimistic results are reported far too often, and useless or harmful treatments become widely used.

Without such enthusiasms about treatments not yet really tested, however, it seems almost certain that we would be worse off in an important way, since we would not have as many good candidates for the randomized trials. Gilbert, McPeek, and Mosteller (1977) have shown that the fraction of randomized trials declaring the novel technique a success seems to be roughly one-half. Perhaps this fraction is about what it ought to be. (If all randomized trials found successes, these trials would be providing only a ritual stamp of approval!).

Expert testimony on legal cases. Often two groups of statisticians acting as expert witnesses will come to very different conclusions by choosing to focus on different aspects of the same data base, or by choosing different techniques.

Modern research in extrasensory perception (ESP). According to its staunchest advocates, ESP has yet to produce a replicable experiment in over 100 years of effort. (See Diaconis (1978) or Jahn (1982, p. 139) for further discussion.) Yet this

flourishing field supports half a dozen journals, several dating back 40 years. What can all the discussion be about?"

There can be no simple answer, but one recurring theme is snooping about in data. A particularly clear example is described by Gardner (1981, Chapter 18), who details some experiments performed by Charles Tart, a parapsychologist at the University of California at Davis. Tart's experiments involved subjects guessing at one of 10 symbols. Some of Tart's subjects guessed correctly far more often than chance predictions. Statisticians looking at Tart's data discovered a faulty random number generator and a fatally flawed protocol. Further, it was not clear just what aspects of the data were to be looked at--in modern ESP testing subjects can get credit for correct guesses that are shifted forward (or back). Clearly replication was called for.

The attempted replication showed no paranormal effect. However, Tart spent many pages in the article that described the replication doing data analysis by new and imaginative methods, testing a variety of hypotheses formed after the data were collected. Sure enough, some of these tests were significant. Most of the parties involved agree that the new tests are merely suggestive, calling for further replication, presumably null results but perhaps some other patterns found, ... Many similar cases of active data snooping can be found in the ESP journal literature.

In summary, things can go wrong in real studies. The psychological work provide a useful list of potential problems. Experiments show that we can be trained to recognize and correct these problems.

1B. MULTIPLICITY--A PERVASIVE PROBLEM

Multiplicity is one of the most prominent difficulties with data analytic procedures. Roughly speaking, if enough different statistics are computed, some of them will be sure to show structure. With the computer, many statistics are forced on us. A few examples will be helpful.

EXAMPLE: PARAPSYCHOLOGY

In a "Ganszfeld" experiment for extrasensory perception (ESP), a "sending" subject concentrates on one of four pictures. A "receiving" subject tries to discern which of the four is being thought of. The four pictures are then given to the receiving subject, who rank orders them--the picture that best fits the impressions received gets rank 1, and the picture that fits worst gets rank 4.

A simple way of testing for ESP is to count the number of times that the picture ranked 1 was the picture actually used. The chance of being correct is 1 in 4, if chance alone is operating. A typical experiment generates 30 permutations. From binomial tables, 12 or more is a significant result at the 5% level.

Other tests could be suggested. For example, one can count the number of times the correct picture was ranked one or two. Another test can be based on the sum of the ranks given to correct pictures (summed over the 30 trials). There are

many other possibilities.

For any single statistic, it is easy to compute a cutoff that would be appropriate if that statistic were used alone. In the Ganszfeld literature cutoffs are chosen so that the chance of a significant result on a single test is 5%. Often, however, many such tests are tried on the same set of 30 rankings. Proponents of ESP have counted a study significant if any one of the tests is significant. This can be very misleading.

Hyman (1982) performed a Monte Carlo experiment to assess the effect of multiple testing. Using the three tests described above (direct hit, top two, sum of ranks), Hyman showed that the chance of at least 1 test significant at 5% is about 15%. Taking account of all the tests actually performed, Hyman concluded that the chance of a significant study was about 25%. Hyman examined all the available Ganszfeld studies and concluded that the proportion of significant studies roughly corresponds to chance expectations. This contrasts sharply with the claims of the proponents of these studies--they compared the proportion of significant studies with 5% and did not allow for multiplicity.

EXAMPLE: DISPLAY PSYCHOLOGY

Problems of multiplicity are likely to be rampant in any really thorough data analysis. For example, in their study of scatterplots, Cleveland, Diaconis, and McGill (1982) computed a difference between two averages for each of 78 subjects in the study. This gave 78 numbers. If the numbers

were "about zero," a certain variable in the study would be judged to have no effect. In analyzing these numbers, many "central values" (e.g., mean, median, biweight) were tried. The final published averages used 10%-trimmed means. For any single average, it is straightforward to derive a significance test. It is less simple, but sometimes feasible, when there are many averages. In this example, replication provided the defense against the problems of multiplicity.

EXAMPLE: LINEAR REGRESSION

Freedman (1983) presents a clear example showing how formal tests can lead us astray if they are applied to data that have been selected as the result of a preliminary analysis. The example involves fitting a linear model. In one simulated case, a matrix was created with 100 rows (data points) and 51 columns (variables). All the entries in this matrix were independent observations drawn from a standard normal distribution, so that, in fact, the columns were independent. The 51st column was taken as the dependent variable Y in a regression equation, and the first 50 columns were taken as the independent variables $X_1, X_2, \ldots, X_{50}$. By construction, Y was independent of the Xs. Indeed the whole matrix consists of "noise." Freedman analyzed these data in two successive multiple linear regressions. The first phase produced the best linear predictor of Y using $X_1, X_2, \ldots, X_{50}$. It turned out that 21 coefficients out of 50 were significant at the 25% level and 5 coefficients out of 50 were significant at

the 5% level. As a crude model of the results of a preliminary analysis, the 21 variables whose coefficients were significant at the 25% level were used in the second stage, which obtained the best linear predictor of Y using these 21 variables. The results were that 13 of the 21 variables were significant at the 5% level. The results from the second pass seem to demonstrate a definite relationship between Y and the X's. That is, between noise and independent noise.

Freedman verifies that the results reported are about what standard theory predicts. In this stark light, the conclusions seem clear: standard statistical procedures do not behave the way the books describe, when applied to data selected as the result of a preliminary analysis. Yet, if we were shown one such trial in which the data had names and a story was woven about the final variables, making their inclusion seem reasonable, many of us might find this conclusion less clear.

Section 10.1 of Leamer (1978) reviews literature on "explaining your results by hindsight." Leamer even introduces a delightfully simple probability model for the kind of selective memory captured by such phrases as "20-20 hindsight" or "Monday morning quarterbacking." Chapter 8 of Leamer (1978) gives some further examples--like the one due to Freedman--which show how much preliminary data screening (or transformation of variables) can distort nominal p-values.

EXAMPLE: MULTIPLE PEEKS AT THE DATA

Multiplicity can be a problem when data come in sequentially, or when a large rich data base is sampled repeatedly or examined in a piecemeal fashion.

For a clear example, consider testing whether the mean of a normal distribution is zero, when the variance is known to be one. The usual 5% test rejects if $|\bar{X}| > 1.96/\sqrt{n}$. Suppose that the n observations are available sequentially and that the usual test is run on the first observation, then again with the first two observations, then again with the first three observations, and so on. The testing stops if any of the tests are significant or if the cutoff of n is reached.

For any single test, the chance of (falsely) rejecting the null hypothesis of zero mean is 5%. When n = 2, the chance that one of two tests rejects goes up to about 8%. When n = 5, the chance is about 13.5%; and when n = 10, the chance is about 19%. Multiple tests increase the chance of a mistake. The numbers used here are based on approximations derived by Siegmund (1977).

Sequential analysts study ways of adjusting standard tests to account for multiple looks. Pocock (1977) shows that, when n = 2, the test that rejects if $|X_1|$ or $\frac{|X_1+X_2|}{\sqrt{2}} > 2.18$ has about 5% chance of error. The appropriate cutoff is 2.42 when n = 5 and 2.56 when n = 10.

Again, put in this stark light, the conclusion seems clear. In a real scientific situation the problem might arise as follows: On a hunch, we look at an average using a small part of our data. Nothing happens, so we look at more data. If the hunch is strong, we may well look again at all the data.

Four examples of multiplicity have been given above. Some other examples are:

- preliminary data selection and screening,
- comparison of many "treatments" either with a standard or with each other,
- transformation of variables.

Not much work has been done on the effect of the first example. The second example is the problem of multiple comparisons. A thorough discussion appears in Miller (1981). Some discussion of the effect of transformation of variables is in Bickel and Doksum (1981).

In Section 1C we review some of the standard cures for these problems.

1C. SOME REMEDIES

Statistical practice has developed a number of useful ways of dealing with the problems posed by exploratory analysis. This section describes some practical remedies. Section 1D describes some more theoretical remedies.

Remedy 1: Publish without P-Values

A number of fine exploratory analyses have been published with a single probability computation or P-value. In publishing such a study, an investigator is distinguishing between confirmatory analysis and exploratory analysis. Ideally, the paper would call attention to the potential subjective elements and might even contain a discussion of how much data snooping went into the final analysis. If confirmatory and exploratory analysis appear in the same document, the two efforts should be clearly distinguished.

Calling attention to the problem causes much of the controversy to go away. Of course, such a study will not be published unless the results are exceptionally informative descriptions of interesting data or are striking and clearly worthy of followup. We now briefly describe some of these studies, both as fine examples and to show that journals do accept interesting data analyses without P-values.

EXAMPLE: AIR POLLUTION

In a series of papers in Science and other journals, Cleveland et al. (1974, 1976a, 1976b, 1979) and Bruntz et al. (1974) have investigated air pollution in Eastern cities in the U.S. Their results are striking. To give one example: ozone

is a secondary pollutant, believed to be produced by two primary pollutants "cooking" in the atmosphere. The primary pollutants are lower on weekends than on weekdays. Cleveland et al. (1974) found that, on average, ozone was slightly higher on weekends. This suggests that we do not yet understand how ozone is produced. These papers are particularly noteworthy because of the notorious difficulty of working with air pollution data. The available data base is huge. At best, pollution data show great spatial and temporal variability. Worse yet, poor methods of measurement and recording often render such data unreliable. Many groups around the country are trying to fit more or less standard statistical models to air pollution data. I think it is fair to say that the Bell Labs group, using exploratory techniques, has triumphed where classical techniques have faltered.

EXAMPLE: ECONOMICS

Economic data often involve many variables, some of them imprecisely measured. Chen, Gnanadesikan and Kettenring (1974) studied a large number of American corporations. Their main purpose was to determine a fair rate of return on investments for American Telephone and Telegraph Corporation by finding other companies that experience "comparable risk." They considered variables thought to be related to risk, such as variability of stock price, debt ratio, and other standard financial variables and looked for companies most like AT&T in the values of these variables. The data base contained 60

variables for each of 10 years on each of 4200 companies. Preliminary considerations reduced these to 14 variables for 10 years on each of about 90 companies. The investigators then attempted to find structure, or clusters, for a single year in the 14-dimensional space of companies. Typical conclusions of this phase of the study were that:

1. AT&T falls in the center of "industrials" but in the extremes of "utilities." This was considered important because in some previous rate cases AT&T had been compared with utilities.
2. The companies most similar to AT&T generally had a much higher rate of return than AT&T.

The investigators then demonstrated that their preliminary conclusions were reasonably stable over the 10 years of available data.

The AT&T study involved a huge data base and wholesale use of large computers. Slater (1974, 1975) uses graphs and other EDA tools to explore economic data on a smaller scale. Leamer (1978) discusses all of econometric model building from a data analysis perspective. His book contains numerous other examples of data analysis in economic settings.

EXAMPLE: MEDICINE

The main conclusions of Reaven and Miller (1979) flow from a single picture. Their study examines the relationship between chemical diabetes and overt diabetes. Diabetes has been considered a homogeneous disorder, caused primarily by the

body's failure to secrete enough insulin. This view suggests that the data should show a smooth continuum ranging from patients with the mildest discernible degree of glucose intolerance through patients who depend upon insulin from outside sources to prevent death. For each of the 145 adult subjects in the study, five variables were measured. The variables, somewhat crudely described, are (1) relative weight, (2) a measure of glucose tolerance, (3) a second measure of glucose tolerance called glucose area, (4) a measure of insulin secretion called insulin area, and (5) a measure of how glucose and insulin interact abbreviated SSPG.

The data were put onto the PRIM-9 graphics system at the Stanford Linear Accelerator Center. The system makes "scatter plots" of 3 variables showing the points on a video screen; when the display is "rotated," the point cloud moves in such a way that points closest to the viewer move in one direction, while those furthest away move in the opposite direction. Parallax fools the eye into seeing three dimensions. The system allows investigators to move the three-dimensional space being viewed into any chosen direction in the data space, which is five-dimensional for the diabetes data.

Figure 1-2 shows an artist's rendition of one of the views found. The picture shows a fat middle and two wings. The middle contains normal patients, and the two wings contain patients with chemical and overt diabetes, respectively. There is no occupied path from one form of diabetes to the other, except through the region occupied by the normal people. This

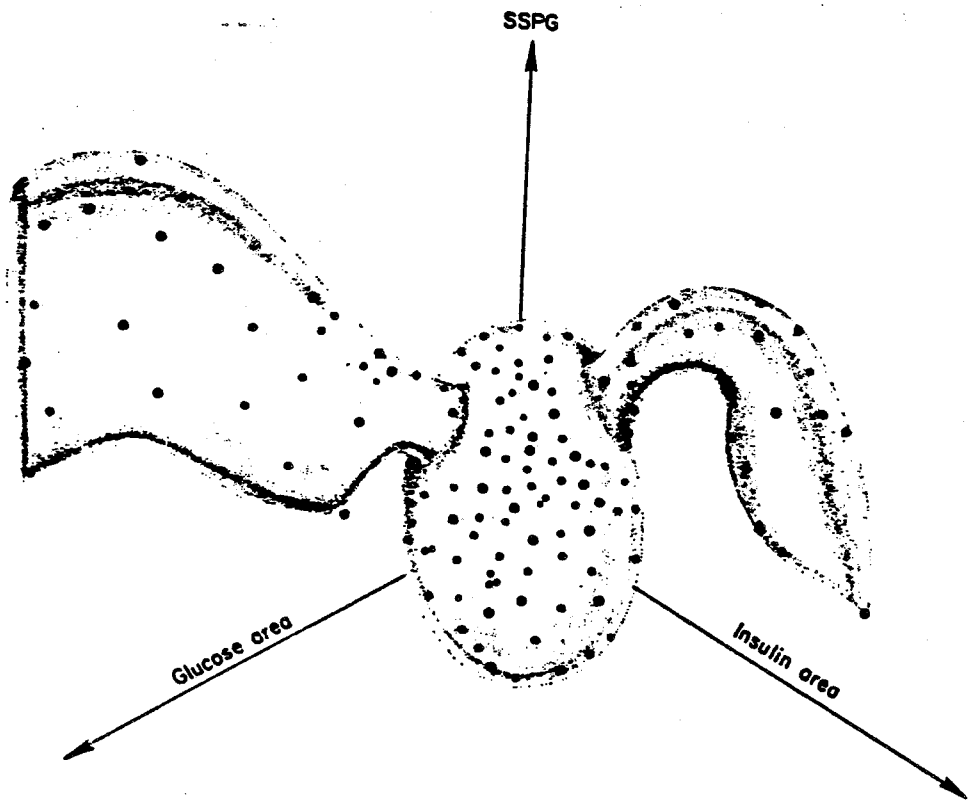
suggests that the usual "smooth transition" model is wrong; indeed, patients with chemical diabetes seldom develop overt diabetes. Reaven and Miller draw further conclusions and suggest followup studies that are now under way. In further graphical analysis of this data set, Diaconis and Friedman (1982) show that a method of drawing four-dimensional scatter plots reveals some additional structure that relates the three variables in Figure 1-2 to relative weight.

Figure 1-2 about here

EXAMPLE: PSYCHOLOGY

Carlsmith and Anderson (1979) give a nice example of exploratory techniques (and common sense) overturning a finding bolstered by P-values. Prevalent folklore suggests that riots tend to occur during periods of very hot weather. A previous study examined 102 major riots in the United States between 1967 and 1971 and concluded that temperature and riots were not

Figure 1-2. Artist's rendition of diabetes data as seen in three dimensions. View is approximately along 45° line as seen through PRIM 9 program on the computer; coordinate axes are in the background.



Source: Reaven and Miller (1979).

monotonically related. Instead, that study concluded that the incidence of riots increases with temperature up to 81-85°F and then decreases sharply with further increases in temperature.

Carlsmith and Anderson challenged this finding. They argued that the previous study had not accounted for the proportion of days during summer in each temperature range. For example, if days in the 81-85°F range are more common than days in the 91-95° temperature range, there are more opportunities for riots in the former range. Carlsmith and Anderson collected fresh data and, using a variety of EDA techniques, demonstrated convincingly that the likelihood of a riot increases monotonically with temperature.

All of these studies sketched above reach their conclusions without using probability. In a survey article, Mallows (1979) describes several other studies with the same features. We conclude that publishing without P-values is often a viable method of communicating an exploratory data analysis.

Remedy 2: Try to Quantify and Correct for the Distortion

There are several techniques for analyzing how a standard statistical procedure behaves on data selected as "interesting" by an exploratory data analysis.

Specific Allowance for Multiplicity

Problems of multiplicity are reviewed in Section 1B. A broad range of techniques for dealing with multiple comparisons are described in Miller (1981). One of the most useful

techniques has turned out to be a simple system of inequalities called the Bonferroni inequalities. These generalize the fact that the probability of a union is smaller than the sum of the individual probabilities; thus the chance that at least one test is significant is smaller than the sum of the chances that individual tests are significant.

Should we correct for multiplicity? After all, we have been practicing data analysis with P-values for many years. If the consequence of the average amount of distortion introduced by skilled workers had been too bad, we can expect that practices would gradually have changed, just as the comparison of the variability of two samples by an F-test appears only in books and papers by innocent authors. This line of thinking suggests a dichotomy.

- For routine analysis in established fields, we want to keep the impact of multiplicity close to where it has traditionally been in the practice of each field's leaders.
- With no tradition, or with multiplicity far in excess of tradition, adjustment seems mandatory.

We suspect that slow modification in the scientific and technological cultures will eventually ensure that techniques are used in somewhere near the optimal manner.

Bayesian Quantification

Many of the problems addressed in this chapter can be approached by using tools developed by Bayesian statisticians.

A book-length review of this literature is given by Leamer (1978), who also presents a theoretical framework for evaluating data analytic studies on non-experimental data. We review this framework in Section 1D.

To illustrate the Bayesian approach to the problems associated with data analysis, we describe a single example: the astronomical relation known as Bode's law. Leamer discusses this in Section 9.5. Bayesian calculations for Bode's law have also been described by Good (1969) and Efron (1971). The following account draws on all three sources.

In 1772, J. E. Bode gave a simple rule for the mean distance of a planet from the sun as a function of the planet order. Let d_n be the distance from the sun to the n^{th} planet from the sun. Bode's law predicts that

$$d_n = 4 + 3\times 2^n \ .$$

For the first eight planets, it gives: 4, 7, 10, 16, 28, 52, 100, 196 (using $n = -\infty, 0, 1, 2, \ldots, 6$). The seven planets known in 1800 had mean distances of 3.9, 7.2, 10, 15.2, 52, 95, 192--the units have been chosen so that the distance of the earth to the sun is 10. The law certainly seems to do well, aside from a missing planet 28 units from the sun. This led a group of astronomers to search the heavens at roughly 28 units from the sun. They found a "planet," actually the asteroid series. Clearly the predictive success of Bode's law adds to its believability.

The obvious question is whether Bode's law is real--in the sense that planets around other stars show a similar geometric relationship. An alternative to "reality" is that Bode's law is a numerological artifact, the result of fooling around with simple progressions. After all, the comparison involves only 8 numbers. A sequence like $a + b \times c^n$ has 3 free parameters. If the planetary distances were approximately in the proportions 1, 2, 5, 10, 17,..., it seems certain that Bode or one of his colleagues would have noticed this and proposed a law of the form $a + b \times n^2$. There are many further possibilities.

Although difficult and different, the problem of testing "the reality of Bode's law" is not intractable. The basic idea is to describe the steps in the data analysis, set up a way of generating data, and see how often a "law" that fits as well can be found. All the steps are difficult, but perhaps not impossible. Quantifying the steps in the data analysis is perhaps the most difficult phase. Still, even thinking of alternatives is a useful exercise. As to generating more data, the simplest approach is to find a new sample (e.g., data from another solar system). Failing this, a simple mathematical generating mechanism may be tried.

Good suggested the following mechanisms for testing B: "Bode's law is true" versus $\bar{B}$: "Bode's law is not true."

1. Under B the logarithms of the planetary distances are independent Gaussian with mean log $(a+b \times 2^n)$ and common variance.

2. Under $\overline{B}$ the logarithms of the planetary distances are like the ordered lengths between points dropped at random into an interval.

Good carried out the testing in a Bayesian framework, putting prior distributions on the parameters involved. He concludes that the odds are 30^4 to 1 in favor of Bode's law. Efron offers a criticism of Good's formulation and various alternatives. He concludes the odds for "the reality of Bode's law" are roughly even. Both authors discuss how similar analyses might be carried out for testing the validity of other simple laws.

The ideas introduced by Good, Efron, and Leamer are fresh and deserve further trials on more down-to-earth examples. They offer one potential route through the difficult maze of quantifying the results of a truly exploratory analysis.

Remedy 3: Try it Out on Fresh Data

Replication on fresh data, preferably by another group of experimenters, is a mainstay of "the scientific method." If done skillfully, replication can eliminate all difficulties with extensive exploration.

Birthday and Deathday

Here is an example of a published, purely data analytic finding that seems nonreplicable. In "Deathday and Birthday" David Phillips (1972) found that people's birthdays and time of death are associated, more people dying just after their birthday than just before. The effect seemed more pronounced among famous people. The evidence consisted of some averages and graphs.

I put Phillips' findings to a test during a course on sample surveys at Stanford University during the winter quarter of 1980. Thirteen students in the course tested Phillips' claim on new data as part of their final project. The students read Phillips' article, each designed a test statistic, and then each took a sample from a different source, such as Who's Who. Without exception, each student's formal test rejected Phillips' hypothesis. Some further analysis of the birthday-deathday problem, again rejecting Phillips' hypothesis, is in Schulz and Bozerman (1980).

While dealing with exploration, replication need not cure other, non-statistical, ills to which science is occasionally subject. Blondlot's N rays (see Klotz (1980) for a modern account) is a well-known case of a widely repeated experiment that turned out to be a complete artifact. Some further examples are described in Chapter 5 of Vogt and Hyman (1979).

As emphasized in the discussion of scientific thinking in the introduction to this section, there are many branches of science but few of technology where "new data" are impossible to obtain.

Remedies 4, 5, and 6 give some forms of partial replication which can be used when new data is hard to come by.

Remedy 4: Borrowing Strength

When direct repetition is not feasible, we often turn to more-or-less parallel situations. The results of their analyses can then be borrowed, to add strength to the result of analyzing the data of our prime concern. Clearly such borrowing has to be done carefully and in the light of the best subject-matter insight available. Equally, it often is the best available way to support--or challenge--the apparent results of our prime analysis. Less obviously, as discussed in Section 1E below, it can play an important part in protecting us from real dangers of a still more subtle kind.

A simple example arises in the analysis of variance, where different groups are assumed to have different average values, but common variability about each group's average. Under this assumption we can use all the data to estimate the common variability--thus borrowing strength--and wind up with more accurate confidence intervals for each group's average.

A second example of borrowing strength comes from the literature on combining tests of significance. Sometimes studies indicate a trend, but the indication is not striking enough to be judged significant. If several such studies show a trend in the same direction, the significance tests can be combined and may result in a significant judgment. A convenient reference to this literature is Rosenthal (1978).

The most important examples of borrowing strength are informal judgments we make all the time, as when we use someone's past performance on a variety of tasks to help judge whether his or her work deserves serious considerations. Tools for quantifying these more subjective aspects can be found in the literature on subjective probability under the label "combination of evidence." Diaconis and Zabell (1982) and Shafer (1976) provide references and discussion.

Remedy 5: Cross-validation

In many problems, such as Bode's law, replication is not a practical possibility. One compromise, which works for large data sets, is what statisticians call cross-validation. The idea is to take a (random) sample from a data set, use exploratory techniques on the sample, and then try the result out on the rest of the data. Mosteller and Tukey (1977) and Efron (1982) contain further discussion. For a careful application in the field of geomagnetism, see MacDonald and Ward (1963).

Remedy 6: Bootstrapping the Exploration

Efron (1979) describes a simple, widely applicable method called the "bootstrap" for assessing variability in a data

analysis. The idea is to use the available sample as a model of the population and draw further samples from the available sample to assess variability. Efron and Gong (1983) and Gong (1982) put the bootstrap to work at the exploratory data analysis level.

In one example a medical investigator had a "multiple discrimination" problem. He had studied 180 patients with a liver disease. For each patient, he measured about 20 variables--age, blood pressure, etc. He also knew whether the patient had lived through the disease or died during the course of the disease. His goal was to investigate the association of the measured variables with the end result. He was seeking a simple rule predicting the outcome from the measured variables.

The investigator first performed preliminary exploratory analysis to get down from 20 measurements per patient to 5. He then used a standard technique called logistic regression to get a rule for predicting outcomes from the five measured variables. It turned out that only three of the five variables really mattered in the prediction rule. The investigator was working at understanding the chemistry of these variables and the effect of changing the levels of some of them in the hope of saving patients.

To investigate the stability of the analysis, Efron and Gong applied the bootstrap. They regarded the set of 180 patients as a population, and they drew a new sample of 180 from this population with replacement. Then, by carefully

quizzing the investigator and his assistants, they replicated the entire data analysis: the preliminary data screening, initial variable selection, logistic discrimination, and final selection of variables. They went through this procedure 500 times. Among many other findings, they noted that no single measured variable appeared in more than half of the final 500 sets of predictor variables. This amount of instability suggests caution in taking any single predictor very seriously.

The exercise just described took a massive amount of computing and months of work. It probably results only in a very crude approximation to the effect of rooting about in the data. Nevertheless, it represents a significant breakthrough in an extremely hard problem.

Remedy 7: Remedies to Come

As data analysis becomes more widely recognized, and used, we can look forward to more research aimed at aiding the analyst in avoiding self-deception. Here are two suggestions. In the middle of an interactive data analysis session it might be useful to have available a display of random, unstructured noise. This should come in a form close to the data being examined. It is easy to imagine structure, dividing lines, and outliers in uniformly distributed scatter plots.

A second suggestion for research is to study adaptations or specializations of the heuristics and biases described by Tversky and Kahneman (1974) along with Nisbett and Ross (1980) to the situations encountered in routine data analysis.

1D. THEORIES FOR DATA ANALYSIS

None of the classical theories of statistics comes close to capturing what a real scientist does when making inferences in a real scientific problem. All formal theories--Neyman-Pearson, decision-theoretic, Bayesian, Fisherian, and others--work with pre-specified probability models. Typically, independent and identically distributed observations from a distribution are supposed known up to a few parameters. In practice, of course, hypotheses are often formed after the data have been examined; patterns seen in the data combine with subject-matter knowledge in a mix that has so far defied description.

This section surveys some alternatives to the more usual theories of inference. These alternative frameworks have been created as a step toward foundations for exploratory data analysis. Two useful categories emerge: theories that explicitly avoid probability and theories that depend on subjective interpretation of probability.

Probability-free Theories

The classical approaches to comparing and interpreting statistical techniques depend on probability. Finch (1979) and Mallows (1983) have started to develop a theory of data description that does not depend on assumptions such as random samples or stochastic errors. They offer a framework for comparing different summaries and methods for assessing how much of the data a given description captures.

An example adapted from Mallows (1983) treats inaccuracy in a non-stochastic setting. We begin with a descriptor δ that takes data sets x in a space X into descriptions in a space D. Informally, the inaccuracy of δ measures how closely the description approximates the data. Formal treatment requires a measure of distance d(x,y) between two data sets x and y. The inaccuracy $i(\delta(x))$ of the descriptor δ of the data set x is defined as the distance between the most distant data sets y,z with the same description:

$$i(\delta(x)) = \max_{y,z}\{d(y,z) \mid \delta(x) = \delta(y) = \delta(z)\}$$

Observe that, with this definition, $i(\delta(x)) = \infty$ means that some data sets have the same description as x and yet are arbitrarily different from x, whereas $i(\delta(x)) = 0$ implies that x is uniquely described by δ.

For example, consider the inaccruacy of a description of the batch $x_1, x_2, \ldots, x_n$ of real numbers. Let $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ denote the ordered values. One reasonable choice of the distance between two batches uses the sum of squares of the differences between the corresponding ordered values:

$$d^2(x,y) = \frac{1}{n}\Sigma(x_{(i)} - y_{(i)})^2 .$$

Freedman and Bickel (1981) discuss some properties of this measure of distance: it is the Mallows metric between the empirical distribution functions of x and y. Table 1-1 gives the squared inaccuracy $i^2(\delta(x))$ for several common descriptors of a batch based on this

distance. The entries in the table are straightforward to derive. For example, the inaccuracy for the median is ∞ because there are data sets with the same median but arbitrarily different order statistics.

Table 1-1 about here

For the endpoints descriptor, the maximum distance i^2 is achieved by choosing one data set with n-1 values equal to $x_{(1)}$ and one value equal to $x_{(n)}$ and the second data set with n-1 values equal to $x_{(n)}$ and one value equal to $x_{(1)}$. The derivation of the inaccuracy for the other descriptors is similar.

Among the summaries based on order statistics the median is least accurate, followed by the 3-number and 5-number summaries, followed by the empirical cumulative distribution function (ECDF). For roughly symmetric batches and large n, the 3-number summary can be seen to be almost 4 times as accurate as the endpoints summary. Under similar assumptions, the 5-number summary is about 4 times as accurate as the 3-number summary.

Table 1.1. Inaccuracy of common batch descriptors--i^2

based on the sum of squares of order statistics

Descriptor δ	Description $\delta(x)$	Inaccuracy $i^2(\delta(x))$
median	Q_2	∞
endpoints	$(x_{(1)}, x_{(n)})$	$(x_{(n)}-x_{(1)})^2 \frac{n-2}{n}$
3-number summary	$(x_{(1)}, Q_2, x_{(n)})$	$\{(x_{(n)}-Q_2)^2 + (Q_2-x_{(1)})^2\} \frac{n-3}{n}$
5-number summary	$(x_{(1)}, Q_1, Q_2, Q_3, x_{(n)})$	$\{(x_{(1)}-Q_1)^2 + (Q_2-Q_1)^2 + (Q_n-Q_2)^2 + (x_{(n)}-Q_3)^2\} \frac{\check{n}-5}{n}$
mean	$\bar{x}$	∞
mean and standard deviation	$(\bar{x}, s)$	$s^2 \frac{2(n-2)}{n}$
ECDF	F_n	0

Another property of a description that does not depend on probability is the breakdown point. Informally, the breakdown point of a description is β if we can change the description an arbitrary amount by changing $n\beta$ data points (but not fewer). Thus the mean has breakdown point 0, the median has breakdown point 1/2, and γ-trimmed means have breakdown point γ. It is possible to develop a reasonable amount of theory from this idea. Some of this is in Mallows (1983), and some of it can be found in Donoho and Huber (1983).

The examples just described are a beginning of a probability-free language for descriptions. Finch (1979) has applied this language to a kind of inference. Working with a finite set Δ of descriptors, he defines the descriptive power of δ at x as the proportion of descriptors in Δ that describe x more accurately than δ does. Finch applies this approach to two-sample problems and 2×2 tables. The ideas are closely related to the nonstochastic interpretation of significance testing developed by Freedman and Lane (1983).

To illustrate, we use data collected as part of an investigation into the possibility of sex bias in graduate

admissions at the University of California--Berkeley. Table 1-2 gives the numbers of men and women accepted and rejected among all applicants for admission to one of Berkeley's largest departments for the 1973-1974 academic year. The acceptance rate for men applicants was 28%. For women, it was 24%. Is the difference between 24% and 28% real? One approach is to compute a chi-squared test for independence. For this table, the significance probability is about 0.26, suggesting that the difference could easily be accidental. The usual interpretation of the chi-squared test depends on assumptions such as random sampling from larger populations. These are at best imprecise here--the data derive from the inherently nonrepeatable graduate admissions process of 1973-74--they are not a sample from any reasonably clearly defined population.

Table 1-2 about here

Freedman and Lane (1983) offer an interpretation for the significance probability 0.26: suppose we divide the pool of applicants into two groups according to whether they are right- or left-handed, or whether their last name begins with A-L or M-Z, or just "at random" (so that no name like "sex" or "handedness" can be associated to the division). For each such division of the applicants into two groups, we could form a table, as in Table 1-2, by counting those admitted and those not admitted in each group. For each table, a chi-squared test can be performed. Freedman and Lane prove that the

Table 1-2. Numbers of men and women accepted and rejected among applicants to a large department at the University of California--Berkeley, for graduate study in academic year 1973-74.

	Accepted	Rejected	Total
men	54	137	191
women	94	299	393
total	148	436	584

Source: D. A. Freedman and D. Lane (1983). "Significance testing in a nonstochastic setting." In P. J. Bickel, K. A. Doksum, and J. L. Hodges, Jr. (Eds.), A Festschrift for Erich L. Lehmann. Belmont, CA: Wadsworth International Group (Table 1, p. 187).

proportion of divisions for which the observed chi-squared statistic is larger than the statistic for the original division by sex is about 0.26.

Their result gives a way to interpret the significance probability 0.26--it is the proportion of divisions showing an inequity at least as large as the original division. Clearly most of these divisions must be regarded as irrelevant to the admissions process. Thus the 4% difference between the proportions of male and female applicants accepted for admission is not unusual. For a fascinating analysis of the full set of data underlying Table 1-2, see Bickel et al. (1975).

If the division into two groups is taken as a description (e.g., percent males, percent left-handed), their approach is similar to the approach of Finch--they count the proportion of descriptions with a more extreme value of the goodness-of-fit criterion. Freedman and Lane also develop a similar rationale for the usual tests in multiple linear regression.

Martin-Lof (1974) develops a model free interpretation for some of the same test statistics based on counting. The motivation and final results are similar to those of Mallows, Finch, Freedman, and Lane. We can look forward to a growing more unified theory as these accounts merge and expand.

Bayesian Theories

Subjective judgments and background knowledge underlie many of the steps taken during an exploratory analysis. The subjective Bayesian approach to statistics seems a natural tool to apply to EDA. In the subjective approach, probability represents the investigator's state of knowledge. Probability changes, after contact with data, via Bayes' theorem.

Connections between Bayesian statistics and EDA are discussed by Box (1980), Dempster (1983), Good (1982), and Leamer (1978). This section discusses some implications of their work. It ends with an explanation of why the Bayesian approach has had little impact on the practice of EDA.

Dempster (1983) suggests that the difference between EDA and statistical modeling is often exaggerated. Although statistical model building makes use of formal probability calculations, the probabilities usually have no sharply defined interpretation (either frequentist or Bayesian), and the whole model building process is simply a form of exploratory analysis. Dempster notes that both EDA and statistical modeling cycle back and forth between fitting curves and looking at residuals. EDA generally starts with summaries and displays, whereas the modeling approach starts by fitting a curve. If the cycling back and forth is carried out diligently and skillfully, the starting place should not have a great effect on the final reported results.

This leads Dempster to suggest fitting truly large models in the exploratory phases--"a model as big as an elephant" as Jimmy Savage is reported to have said. Such a model might well have as many parameters as data points; and Bayesian methods represent a possible way to make parameters identifiable. Dempster goes on to compare the approach just outlined (using probability) with non-probabilistic approaches.

Good (1982) does not attempt to present a full-blown theory of data analysis. He describes some of the things that such a

theory must cover (Mallows and Walley (1981) do this provocatively as well). He emphasizes that such a theory will involve psychology, in that techniques of description and display must be designed to match salient features of a data set to human cognitive abilities. For Good, EDA is concerned mainly with the encouragement of hypothesis formulation. Good discusses the problem of distinguishing a pattern in the data from a mere coincidence. He argues that it is possible to judge that a pattern has an underlying explanation even if we are unable to specify the explanation sharply. This suggests that patterns with extremely small prior probability of being potentially explicable (given the context) will be discarded. Patterns with an appreciable prior probability of being potentially incorporated in a useful hypothesis for explaining the data will be displayed and labeled "salient features of the data."

One quantification of "explicativity" can be based on subjective probability. For example, if H is a hypothesis about the data ("this drug really has an effect" or "the experimenter was cheating"), then $P(H|G)$ denotes the investigator's prior probability of H based only on the background information G. If E is an event or pattern ("80 of 100 patients recovered" or "the successful trials are at prime numbers"), then $P(E|H,G)$ denotes the likelihood that E occurs, given G and H.

Informally, a good explanation of an event or pattern E is a hypothesis H such that $P(E|H,G) >> P(E|G)$ but $P(H|G)$ is

not too small. Good offers a more precise quantification of explicativity as

$$\log P(E|H,G) - \log P(E|G) + \frac{1}{2} \log P(H|G) .$$

The difference between the first two terms measures how much H increases the likelihood of E. The final log term will be a large negative number, thus decreasing explicativity, if the prior probability of H is very small. Although the factor 1/2 could properly be replaced by any other small constant, Good gives an argument to justify it.

Such numerical quantification, however crude, seems mandatory if data analysis is to be automated.

Good's reflections lead to a number of useful practical suggestions. He considers techniques for reducing dimensionality--such as principal components, factor analysis, and multidimensional scaling; these have generally been used to represent high-dimensional data in two dimensions. Good suggests using these techniques to reduce to 6 or 7 dimensions and then using devices like projection pursuit or colored graphics to explore the reduced version of the data.

The book by Leamer (1978) represents an attempt to deal systematically with the problems encountered in exploratory analysis. The subtitle of the book--"ad hoc inference with non-experimental data"--suggests the book's two themes. First, the book attempts to deal with the "data mining, fishing, number

crunching" side of statistics. Second, the techniques are intended for application to data collected in a wide variety of ways, not necessarily as part of a designed, randomized trial. Leamer realizes that the classical version of Bayesian statistics, which requires fully specified probabilistic assumptions, is not particularly relevant to guiding an exploratory analysis. He suggests that the Bayesian approach

is sufficiently flexible to allow suitable iteration, permitting the main ingredients of an exploratory analysis to be legitimate or at least understandable.

Most of Leamer's examples involve regression. Topics covered in detail include "data selection searches," in which the data in hand are used to detect and investigate deviations from initial models (or simple descriptions); "post data model construction," in which models, or theory, are worked up after seeing data; and "simplification searches," in which a larger model is "pruned down" to allow findings to be summarized and communicated. Many of the problems are illustrated on econometric data. Leamer is quite modest about the accomplishments of the theory he sets forth. Nevertheless he has constructed a reasonably complete framework to model the meanderings of a real data analyst. It offers a baseline for comparison and improvement.

To conclude this section, we consider two challenges to Bayesian techniques from EDA. First, exploration is important in problems where nothing is unknown. Consider a complete list of a population--like the records of all employees in a company or a list of features of the world's major rivers. Here nothing is unknown or random--a complete, accurate list is assumed available. Still, a great deal of useful data analytic work is possible in the form of simple descriptions and summaries. Because there are no unknown quantities (or parameters), no guidance for the data

analytic investigation can be expected from the usual models of Bayesian (or frequentist) statistics.

To discuss the second challenge, let us review the subjective Bayesian approach to inference. First, probability represents an investigator's state of knowledge. Second, probability changes, after contact with data, via Bayes' theorem. Tukey (1972, pp. 61-63), Mallows (1970), and Savage (1970) have discussed the limitations of this approach. Here are two of Tukey's objections: "... the discovery of the irrelevance of past knowledge to the data before us can be one of the great triumphs of science." Further, "Bayesian techniques assume we know all the alternate possible states of nature."

Both of these objections seem cogent. They link up with recent work in Bayesian statistics that suggests that Bayes' theorem is not the only way that probability shifts in the light of new evidence. In surveying this recent work, Diaconis and Zabell (1982) describe several other methods of changing an initial probability assessment. In addition to Bayes' rule and complete quantification, Jeffreys' rule--a generalization of Bayes' rule--permits part of the previously quantified probability to be retained. Bayes' rule will most often be the route to probability change in situations where a lot of experience has dictated the relevant variables. In exploratory situations such experience is usually not available. The impact of new data may well be to remind us of numerous variables not previously contemplated. Then Bayes' rule does not apply, and one of the alternative methods of changing probability

must be used. Faced with many complex shifts in probability, the Bayesian data analyst may well decide to abandon the usual Bayesian machinery until the situation has stabilized enough to make such calculations cost-effective.

1E. USES FOR MATHEMATICS

Most of the discussion in this chapter has pointed to the difference between exploratory techniques and the techniques of probability and mathematical statistics. This section shows how the tools developed for classical statistics can lead to insight in EDA. Of course, much of the material in UREDA and in other chapters of this volume has this emphasis.

An important teaching of EDA is to try out lots of summaries. This process necessarily leads to a wealth of techniques. Many of these techniques will not hold much interest outside the specific context for which they were created. This leads some to say that mathematics is not useful for EDA (because the formative, data specific stages are the heart of EDA).

As shown in UREDA and this volume, many exploratory techniques are broadly useful and merit closer study. Mathematical methods can be useful at this stage. Mathematics can help in fine tuning, finding pitfalls, and choosing the better of several exploratory devices.

Careful studies of EDA techniques have made heavy use of the Monte Carlo technique in the spirit of "try it out in a problem where we know the answer and see what happens." Experimental sampling on a computer is a mathematical activity, whether it involves simple repetition or a sophisticated combination of importance sampling and other variance-reduction techniques.

Computational mathematics often tells us "less about more," in that the results are specific and approximate, though

we have great flexibility as to what we may study. Theorem-proving mathematics usually tells us "more about less," in that we come to exact results, or remarkable approximations, but often with severe restrictions on what they apply to (populations may have to be Gaussian or sample sizes may have to be unrealistically large).

The two approaches can be combined with great effect. It seems mandatory to check theorems by computational methods to see whether their conclusions hold in cases of practical interest. Similarly, simulation studies, even if the examples span a broad range, are limited by the imagination of the investigator. It seems mandatory to have some theoretical backup to help organize and verify the computational results. (Many a Monte Carlo has all numbers "off" through a programming error.) The following example shows how theorem-proving mathematics can add insight to a technique that has been very thoroughly studied by simulation.

Tukey's biweight (Chapter 11 of UREDA) is a popular robust estimator of the center of a batch of data. The estimator works with data standardized by a "tuning constant" times a scale estimate (usually the MAD--the median of the absolute deviations about the median). A natural question is: which tuning constants are good, or does it matter? Monte Carlo work on this question proceeds by trying the estimator out in situations where the correct answer and form of error terms are known; the tuning constant is chosen so that the estimator

behaves well in these situations. Based on such studies, tuning constants between 4.8 and 9.1 are in use as of this writing.

Freedman and Diaconis (1982) asked a sharp mathematical question of the biweight: is it consistent? That is, suppose a large sample is taken from a population that is symmetric about zero (say). Suppose too that the population density is positive and smooth about zero. Compute the biweight estimate of the center of symmetry. Will this estimate be close to zero? Freedman and Diaconis show that, for constants smaller than 5.4, the answer is no. For constants larger than 5.4, the biweight is consistent. For most location estimators, like the median or trimmed means, the answer is yes. For the biweight, the answer depends on the tuning constant. The results generalize: in the language of Chapter 11 of UREDA, any M-estimator based on a redescending ψ-function can be inconsistent.

Consistency and inconsistency are tied to increasing sample size. The practical statistician may wonder what these results say about sample size 20 or 100. A rule of thumb is that, if a procedure behaves badly in large samples, it will have similar, objectionable features for moderate sample sizes. In the case of the biweight, the moderate-sample phenomenon is that, for small tuning constants, the biweight is quite tricky to compute numerically. It may oscillate among several answers for a given data set; a slight change in numerical technique can produce a different answer.

The Monte Carlo investigation of the biweight has been going on for about 10 years. It ranks high as a model of thoroughness for the computational approach. Nonetheless, the theorem-proving investigation pointed to a new direction: the counterexamples to consistency are symmetric densities that are multimodal. None of the numerical trials has yet considered such examples. Presumably such cases are rare in practice.

The moral is clear--"In union there is strength." We need to use both computational mathematics and theorem-proving mathematics to guide the growth of our new techniques and insights. This conclusion runs through much of the best statistical practice. Indeed, Student's first derivation of a distribution for Student's t combined experimental sampling with the method of moments to fit one of Karl Pearson's frequency curves. Nearly a decade elapsed before R. A. Fisher provided a formal proof. Empirical and theoretical work have continued through the years. Their union provides a very thorough understanding of the usefulness and validity of Student's t.

Many other novel techniques have yielded to mathematical analysis. In addition to robust techniques and other examples in EDA, the bootstrap (Efron (1979), Bickel and Freedman (1981), cross validation (Efron (1982), projection pursuit (Friedman and Tukey (1974), Diaconis and Freedman (1983)), and recursive partitioning (Breiman et al. (1983)) are worth mentioning. Each of these

began as a novel technique which did not fit neatly within standard statistics. As the techniques became more widely used, they have been successfully analyzed using both kinds of mathematics.

1F. IN DEFENSE OF CONTROLLED MAGICAL THINKING

Magical thinking couples the inclination to seek meaningful connections and interpretations with the disinclination to learn from experience. Our task is to keep the first and control the second.

For the first, the scientist has an obligation to explore--to seek meaningful connections. Edison tried oh so many candidates for a filament to provide the first incandescent lamp. A critical part of the discovery of the continental drift was staring at a map and noticing how the continents fit together. In both cases, exploration was vital, and formal significance was absent or unnoticed. In many areas, progress can only come from exploration.

The second aspect of magical thinking can also be of enormous use: In order to make progress in a complex situation we often assume a model, paradigm, or working hypothesis and proceed as if it were true. This often involves ignoring data and competing theories. Thomas Kuhn (1970, Chapter 2; 1977, Chapter 9) argues that many of the discoveries of science were made because of the unswerving belief in a working hypothesis. Such beliefs are rewarded when

- the belief allows us to forget about philosophical controversy and get down to "honest work" which will be interpretable from most philosophical perspectives
- the beliefs are longshots that prove to be correct
- the beliefs are wrong, but to defend them, we undertake

a detailed series of experiments that ultimately lead to a convincing refutation.

Perhaps Francis Bacon, quoted by Kuhn (1970, p. 18), put it best when he wrote

"Truth emerges more readily from error than from confusion."

There is not much available theory to guide our decisions on how long to cling to a belief unsupported by data. Kuhn makes a clear case for the eminent practicality of this version of magical thinking and documents how revolutions, great and small, occur when such beliefs cease. Let us consider the desire and skill for exploration together with an awareness of the uses and pitfalls of working hypotheses as a controlled form of magical thinking.

Where do we stand? The new exploratory techniques seem to be a mandatory supplement to more classical statistical procedures tied to Gaussian error models and linearity. The argument is two-pronged.

First, exploratory techniques find useful structure where classical techniques fall flat. This is shown in the examples described in remedy 1 of Section 1C. Second, in many situations, the "usual assumptions" are not even approximately valid. Therefore, the resulting P-values or levels of significance do not mean what they say.

Sometimes we may be able to supply more appropriate p-values, as indicated in remedies 2 and 4 of Section 1C. None of the special theories outlined in Section 1D are ready for routine use. The computer and mathematics can combine to give a reasonable way of checking out some more widely used procedures, but again under stringent restrictions.

Second, despite its lack of sophistication, a purely data analytic approach has something to offer. It is an empirical fact that using the tools of EDA to root about in large rich collections of data leads to useful results. One need not be a subject matter expert or be able to explain the patterns in order to proceed usefully.

Of course, we try hard to control the tendency to see patterns in noise and label our findings as exploratory. Despite the cost of false leads, it seems that controlled magical thinking is here to stay.

REFERENCES

Barber, T. X. (1976). Pitfalls in Human Research: Ten Pivotal Points. New York: Pergamon Press.

Bickel, P. J. and Doksum, K. A. (1981). "An analysis of transformations revisited," Journal of the American Statistical Association, 76, 296-311.

Bickel, P. J. and Freedman, D. A. (1981). "Some asymptotic theory for the bootstrap," Annals of Statistics, 9, 1196-1217.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). "Sex bias in graduate admissions: data from Berkeley," Science, 187, 398-404. Reprinted in W. B. Fairley and F. Mosteller (eds.), Statistics and Public Policy, 1977. Reading, MA: Addison-Wesley, pp. 113-130.

Box, G. E. P. (1980). "Sampling and Bayes" inference in scientific modeling and robustness," with Discussion, Journal of the Royal Statistical Society, Series A, 143, 383-430.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1983). Classification and Regression Trees. Belmont, CA: Wadsworth.

Bruntz, S. M., Cleveland, W. S., Graedel, T.E., Kleiner, B., and Warner, J. L. (1974). "Ozone concentration in New Jersey and New York: statistical association with related variables," Science, 186, 257-259.

Bunker, J. P., Barnes, B. A., and Mosteller, F., (eds.) (1977). Costs, Risks, and Benefits of Surgery. New York: Oxford University Press.

Carlsmith, J. M. and Anderson, C. A. (1979). "Ambient temperature and occurrence of collective violence: a new analysis," Journal of Personal and Social Psychology, 37, 337-344.

Cassirer, E. (1972). An Essay on Man. Yale University Press: New Haven.

Chen, H-J, Gnanadesikan, R., and Kettenring, J. R. (1974). "Statistical methods for grouping corporations," Sankhyā, Series B, 36m 1-28.

Cleveland, W. S., Diaconis, P., and McGill, R. (1982). "Variables on scatterplots look more highly correlated when the scales are increased," Science, 216, 1138-1141.

Cleveland, W. S., Kleiner, B., McRae, J. E., and Warner, J. L. (1976). "Photochemical air pollution: transport from the New York City area into Connecticut and Massachusetts," Science, 191, 179-181.

Cleveland, W. S., Kleiner, B., and Warner, J. L. (1976). "Robust statistical methods and photochemical air pollution data," Journal of the Air Pollution Control Association, 26, 36-38.

Cleveland, W. S. and Graedel, T. E. (1979). "Photochemical air pollution in the Northeast United States," Science, 204, 1273-1278.

Cleveland, W. S., Graedel, T. E., Kleiner, B., and Warner, J. L. (1974). "Sunday and workday variations in photochemical air pollutants in New Jersey and New York," Science, 186, 1037-1038.

Dempster, A. P. (1983). "Purposes and limitations of data analysis." In G. E. P. Box, T. Leonard, and C-F Wu (eds.). New York: Academic Press. pp. 117-133.

Diaconis, P. (1978). "Statistical problems in ESP research," Science, 201, 131-136; 1145-1146.

Diaconis, P. (1981). "Magical thinking in the analysis of scientific data," Annals of the New York Academy of Sciences, 364, 236-244.

Diaconis, P. and Freedman, D. A. (1983). "Some asymptotics for graphical projection pursuit," to appear Annals of Statistics, Stanford Technical Report 195.

Diaconis, P. and Friedman, J. H. (1983). "M and N plots," In H. Rizvi, J. Rustagi, and D. Siegmund (eds.), A festschrift for Herman Chernoff. New York: Wiley.

Diaconis, P. and Zabell, S. L. (1982). "Updating subjective probability," Journal of the American Statistical Association, 77, 822-830.

Donoho, D. L. and Huber, P. J. (1983). "The notion of breakdown point." In P. J. Bickel, K. A. Doksum, and J. L. Hodges, Jr. (eds.), A Festschrift for Erich L. Lehmann. In Honor of His Sixty-Fifth Birthday. Belmont, CA: Wadsworth International Group, pp. 157-184.

Efron, B. (1979). "Bootstrap methods: another look at the jackknife," Annals of Statistics, 7, 1-26.

Efron, B. (1971). "Does an observed sequence of numbers follow a simple rule? (Another look at Bode's Law)," Journal of the American Statistical Association, 66, 552-559.

Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans. Philadelphia: SIAM

Efron, B. and Gong, G. (1983). "A leisurely look at the bootstrap, the Jackknife, and cross-validation," The American Statistician, 37, No. 1, 36-48.

Finch, P. D. (1979). "Description and analogy in the practive of statistics," Biometrika, 66, 195-208. (Includes Comments by Others)

Freedman, D. A. (1983). "A note on screening regression equations," The American Statistician, 37, 152-155.

Freedman, D. A. and Diaconis, P. (1983). "On inconsistent M-estimators," Annals of Statistics, 10, 454-461.

Freedman, D. A. and Lane, D. (1983). "Significance testing in a nonstochastic setting." In P. J. Bickel, K. A. Doksum, and J. L. Hodges, Jr. (eds.). In Honor of His Sixty-Fifty Birthday

A Festschrift for Erich L. Lehmann. Belmont, CA: Wadsworth International Group, pp. 185-208.

Friedman, J. H. and Tukey, J. W. (1974). "A projection pursuit algorithm for exploratory data analysis," IEEE Transactions on Computers, C-23, 881-890.

Gabbe, J. D., Wilk, M. B., and Brown, W. L. (1967). "Statistical analysis and modeling of the high-energy proton data from the Telstar satellite," Bell System Technical Journal, 46, 1301-1450.

Gardner, M. (1981). Science: Good, Bad, and Bogus. New York: Prometheus.

Gilbert, J. P., McPeek, B., and Mosteller, F. (1977a). "How frequently do innovations succeed in surgery and anesthesia?" In J. M. Tanur, F. Mosteller, W. H. Kruskal, R. F. Link, R. S. Pieters, G. R. Rising, and E. L. Lehmann, (eds.), Statistics: A Guide to the Biological and Health Sciences. San Francisco: Holden-Day, pp. 51-64.

Gilbert, J. P., McPeek, B., and Mosteller, F. (1977b). "Progress in surgery and anesthesia: benefits and risks of innovative therapy." In J. P. Bunker, B. A. Barnes, and F. Mosteller (eds.), Costs, Risks, and Benefits of Surgery. New York: Oxford University Press, pp. 124-169.

Gnanadesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. New York: Wiley.

Gong, G. (1982). Cross-validation, the jackknife, and the bootstrap: excess error estimation in forward logistic regression. Ph.D. Dissertation, Department of Statistics - Stanford, University.

Good, I. J. (1969). "A subjective evaluation of Bode's Law and an 'objective' test for approximate numerical rationality," with discussions, Journal of the American Statistical Association, 64, 23-66.

Good, I. J. (1982). The Philosophy of Exploratory Data Analysis. Technical Report, Department of Statistics - Virginia Polytechnic Institute.

Hoaglin. D. C. (1977). "Direct approximations for chi-squared percentage points," Journal of the American Statistical Association, 72, 508-515.

Hyman, R. (1982). Does the Ganzfeld experiment answer the critics objection. Paper presented at the combined meeting of the Parapsychology Association and the Society for Physical Research Cambridge, England. To appear Journal of Parapsychology

John, R. G. (1982). "The persistent paradox of psychic phenomena: an engineering perspective," Proceedings of the IEEE, 70, 136-170.

Kahneman, D., Slovic, P., and Tversky, A., (eds.) (1982). Judgment under Uncertainty: Heuristics and Biases. Cambridge: Cambridge University Press.

Klotz, I. M. (May 1980). "The N-ray affair," Scientific American, 242, No. 5, 168-175.

Kuhn, T. S. (1970). The Structure of Scientific Revolutions, Second edition. Chicago; University of Chicago Press.

Kuhn, T. S. (1977). The Essential Tension: Selected Studies in Scientific Tradition and Change. Chicago: University of Chicago Press.

Lakatos, I. (1976). Proofs and Refutations: The Logic of Mathematical Discovery, edited by J. Worrall and E. Zahar. Cambridge: Cambridge University Press.

Leamer, E. E. (1978). Specification Searches: Ad Hoc Inference with Nonexperimental Data. New York: Wiley.

MacDonald, N. J. and Ward, F. (1963). "The prediction of geomagnetic disturbance indices. 1. The elimination of internally predictable variations," Journal of Geophysical Research, 68, 3351-3373.

Mallows, C. L. (1970). "Some comments on Bayesian methods." In D. L. Meyer and R. O. Collier, Jr. (eds.), Bayesian Statistics. Itasca, IL: Peacock, pp. 71-84.

Mallows, C. L. (1970). "Robust methods--some examples of their use," The American Statistician, 33, 179-184.

Mallows, C. L. (1983). "Data description." In G. E. P. Box, T. Leonard, and C-F. Wu (eds.), Scientific Inference, Data Analysis, and Robustness. New York: Academic Press, pp. 135-151.

Mallows, C. L. and Walley, P. (1981). "A theory of data analysis." In 1980 Proceedings of the Business and Economic Statistics Section. Washington, D.C.: American Statistical Association, pp. 8-14.

Miller, R. G., Jr. (1977). "Developments in multiple comparisons, 1966-1976," Journal of the American Statistical Association, 72, 779-788.

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

Mosteller, F. and Tukey, J. W. (1977). Data Analysis and Regression: A Second Course in Statistics. Reading, MA: Addison-Wesley.

Nisbett, R. and Ross, L. (1980). Human Inference: Strategies and Shortcomings of Social Judgment. Englewood Cliffs, NJ: Prentice-Hall.

Phillips, D. P. (1978). "Deathday and birthday: an unexpected connection." In J. M. Tanur, F. Mosteller, W. H. Kruskal, R. F. Link, R. S. Pieters, G. R. Rising, and E. L. Lehmann, (eds.), Statistics: A Guide to the Unknown, second edition. San Francisco, CA: Holden-Day. pp. 71-85.

Pocock, S. J. (1977). "Group sequential methods in the design and analysis of clinical trials," Biometrika, 64, 191-199.

Reaven, G. and Miller, R. (1979). "An attempt to define the nature of chemical diabetes using multidimensional analyses," Diabetologia, 16, 17-24.

Rosenthal, R. (1978). Combining results of independent studies. Psychology Bulletin, 85, 185-193.

Rosenthal, R. (1981). "Pavlov's mice, Pfungst's horse, and Pygmalion's Pons: Some models for the study of interpersonal expectancy effects," Annals of the New York Academy of Sciences, 364, 182-198.

Ross, L. and Lepper, M. R. (1980). The perseverance of beliefs; Empirical and normative considerations. In R. A. Shweder (ed.), New Directions for Methodology of Behaviorial Sciences: Fallible judgment in Behaviorial Research. San Francisco, CA: Jossey Bass.

Savage, L. J. (1970). "The shifting foundations of statistics." In Logic, Laws and Life, edited by R. Colodny. Pittsburgh: University of Pittsburgh Press, pp. 3-18. Reprinted in The Writings of Leonard Jimmie Savage--A Memorial Selection, published by The American Statistical Association and The Institute of Mathematical Statistics, 1981, pp. Washington, D.C.: The American Statistical Association.

Schulz, R. and Bozerman, M. (1980). "Ceremonial occasions and mortality: A second look," American Psychologist, 35, 253-261.

Shafer, G. (1976). A Mathematical Theory of Evidence. Princeton, N.J.: Princeton University Press.

Shweder, R. A. (1977). "Likeness and likelihood in everyday thought: magical thinking in judgments about personality," Current Anthropology, 18, 637-658.

Siegmund, D. (1977). "Repeated significance tests for a normal mean," Biometrika, 64, 177-189.

Slater, P. B. (1974). "Exploratory analyses of trip distribution data," Journal of Regional Science, 14, 377

Slater, P. B. (1975). "Petroleum trade in 1970: An exploratory analysis," IEEE Transactions on Systems, Man, and Cybernetics, SMC-5, 278-283.

Tukey, J. W. (1969). "Analyzing data: sanctification or detective work," American Psychologist, 24, 83-91.

Tukey, J. W. (1972). "Data analysis, computation, and mathematics," Quarterly of Applied Mathematics, 30, 51-65.

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

Tversky, A. and Kahneman, D. (1974). "Judgment under uncertainty: heuristics and biases," Science, 185, 1124-1131.

Vogt, E. Z. and Hyman, R. (1979). Water Witching U.S.A., second edition. Chicago: University of Chicago Press.

Worsley, P. (1960). The Trumpet Shall Sound: A Study of "Cargo Cults" in Melanesia, second edition. New York: Schicken.