Preface

This book is based on an important principle:

It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it.

Learning first what you can do will help you to work more easily and effectively

This book is about exploratory data analysis, about looking at data to see what it seems to say It concentrates on simple arithmetic and easy-to-draw pictures. It regards whatever appearances we have recognized as partial descriptions, and tries to look beneath them for new insights. Its concern is with appearance, not with confirmation.

Examples, NOT case histories

The book does not exist to make the case that exploratory data analysis is useful. Rather it exists to expose its readers and users to a considerable variety of techniques for looking more effectively at one's data. The examples are not intended to be complete case histories. Rather they show isolated techniques in action on real data. The emphasis is on general techniques, rather than specific problems.

A basic problem about any body of data is to make it more easily and effectively handleable by minds--our minds, her mind, his mind. To this general end:

- anything that makes a simpler description possible makes the description more easily handleable.
- anything that looks below the previously described surface makes the description more effective.

So we shall always be glad (a) to simplify description and (b) to describe one layer deeper

In particular:

- to be able to say that we looked one layer deeper, and found nothing, is a definite step forward--though not as far as to be able to say that we looked deeper and found thus-and-such.
- to be able to say that "if we change our point of view in the following way things are simpler" is always a gain--though not quite as much as to be able to say "if we don't bother to change our point of view (some other) things are equally simple"

Thus, for example, we regard learning that log pressure is almost a straight line in the negative reciprocal of absolute temperature as a real gain, as compared to saying that pressure increases with temperature at an evergrowing rate. Equally, we regard being able to say that a batch of values is roughly symmetrically distributed on a log scale as much better than to say that the raw values have a very skew distribution.

In rating ease of description, after almost any reasonable change of point of view, as very important, we are essentially asserting a belief in quantitative knowledge--a belief that most of the key questions in our world sooner or later demand answers to "by how much?" rather than merely to 'in which direction?"

Consistent with this view, we believe, is a clear demand that pictures based on exploration of data should *force* their messages upon us. Pictures that emphasize what we already know--"security blankets" to reassure us--are frequently not worth the space they take. Pictures that have to be gone over with a reading glass to see the main point are wasteful of time and inadequate of effect. **The greatest value of a picture** is when it *forces* us to notice **what we never expected to see.**

We shall not try to say why specific techniques are the ones to use. Besides pressures of space and time, there are specific reasons for this. Many of the techniques are less than ten years old in their present form--some will improve noticeably. And where a technique is very good, it is not at all certain that we yet know why it is.

We have tried to use consistent techniques wherever this seemed reasonable, and have not worried where it didn't. Apparent consistency speeds learning and remembering, but ought not to be allowed to outweigh noticeable differences in performance.

In summary, then, we:

- leave most interpretations of results to those who are experts in the subject-matter field involved.
- present techniques, not case histories.
- regard simple descriptions as good in themselves.
- feel free to ask for changes in point of view in order to gain such simplicity
- demand impact from our pictures.
- regard every description (always incomplete!) as something to be lifted off and looked under (mainly by using residuals).
- regard consistency from one technique to another as desirable, not essential.

Confirmation

The principles and procedures of what we call confirmatory data analysis are both widely used and one of the great intellectual products of our century

In their simplest form, these principles and procedures look at a sample--and at what that sample has told us about the population from which it came--and assess the precision with which our inference from sample to population is made. We can no longer get along without confirmatory data anlysis. **But we need not start with it.**

The best way to **understand what CAN be done is no longer**--if it ever was--**to ask what things could,** in the current state of our skill techniques, **be confirmed** (positively or negatively). Even more understanding is *lost* if we consider each thing we can do to data *only* in terms of some set of very restrictive assumptions under which that thing is best possible--assumptions we *know we CANNOT check in practice*

Exploration AND confirmation

Once upon a time, statisticians only explored. Then they learned to confirm exactly--to confirm a few things exactly, each under very specific circumstances. As they emphasized exact confirmation, their techniques inevitably became less flexible. The connection of the most used techniques with past insights was weakened. Anything to which a confirmatory procedure was not explicitly attached was decried as "mere descriptive statistics", no matter how much we had learned from it.

Today, the flexibility of (approximate) confirmation by the jackknife makes it relatively easy to ask, for almost any clearly specified exploration, "How far is it confirmed?"

Today, exploratory and confirmatory can--and should--proceed side by side. This book, of course, considers only exploratory techniques, leaving confirmatory techniques to other accounts.

Relation to the preliminary edition

The preliminary edition of *Exploratory Data Analysis* appeared in three volumes, represented the results of teaching and modifying three earlier versions, and had limited circulation. Complete restructuring and revision was followed by further major changes after the use of the structure and much of the material in an American Statistical Association short course. The present volume contains:

- those techniques from the first preliminary volume that seemed to deserve careful attention.
- a selection of techniques from the second preliminary volume.
- a few techniques from the third preliminary volume.
- some techniques (especially in chapters 7, 8, and 17) that did not apppear in the preliminary edition at all.

It is to be hoped that the preliminary edition will reappear in microfiche form.

About the problems

The teacher needs to be careful about assigning problems. Not too many, please. They are likely to take longer than you think. The number supplied is to accommodate diversity of interest, not to keep everyone busy

Besides the length of our problems, both teacher and student need to realize that many problems do not have a single "right answer" There can be many ways to approach a body of data. Not all are equally good. For some bodies of data this may be clear but for others we may not be able to tell from a single body of data which approach is preferred. Even several bodies of data about very similar situations may not be enough to show which approach should be preferred. Accordingly, it will often be quite reasonable for different analysts to reach somewhat different analyses.

Yet more--to unlock the anlysis of a body of data, to find the good way or ways to approach it, may require a key, whose finding is a creative act. Not everyone can be expected to create the key to any one situation. And, to continue to paraphrase Barnum, no one can be expected to create a key to each situation he or she meets.

To learn about data anlysis, it is right that each of us try many things that do not work--that we tackle more problems than we make expert analyses of. We often learn less from an expertly done analysis than from one where, by not trying something, we missed--at least until we were told about it--an opportunity to learn more. Each teacher needs to recognize this in grading and commenting on problems.

Scratching down numbers (stem-and-leaf) 1

chapter index on next page

1A. Quantitative detective work

Exploratory data analysis is detective work--numerical detective work--or counting detective work--or graphical detective work.

A detective investigating a crime needs both tools and understanding. If he has no fingerprint powder, he will fail to find fingerprints on most surfaces. If he does not understand where the criminal is likely to have put his fingers, he will not look in the right places. Equally, the analyst of data needs both tools and understanding. It is the purpose of this book to provide some of each.

Time will keep us from learning about many tools--we shall try to look at a few of the most general and powerful among the simple ones. **We do not guarantee to introduce you to the "best" tools, particularly since we are not sure that there can be unique bests.**

Understanding has different limitations. As many detective stories have made clear, one needs quite different sorts of detailed understanding to detect criminals in London's slums, in a remote Welsh village, among Parisian aristocrats, in the cattle-raising west, or in the Australian outback. We do not expect a Scotland Yard officer to do well trailing cattle thieves, or a Texas ranger to be effective in the heart of Birmingham. Equally, very different detailed understandings are needed if we are to be highly effective in dealing with data concerning earthquakes, data concerning techniques of chemical manufacturing, data concerning the sizes and profits of firms in a service industry, data concerning human hearing, data concerning suicide rates, data concerning population growth, data concerning fossil dinosaurs, data concerning the genetics of fruit flies, or data concerning the latest exploits in molecular biology. A full introduction to data analysis in any one of these fields--or in any of many others--would take much more time than we have.

The Scotland Yard detective, however, would be far from useless in the wild west or the outback. He has certain general understandings of conventional detective work that will help him anywhere.

In data analysis there are similar general understandings. We can hope to lead you to a few of them. We shall try.

The processes of criminal justice are clearly divided between the search for the evidence--in Anglo-Saxon lands the responsibility of the police and other

investigative forces--and the evaluation of the evidence's strength--a matter for juries and judges. In data analysis a similar distinction is helpful. Exploratory data analysis is detective in character. Confirmatory data analysis is judicial or quasi-judicial in character. Only exploratory data analysis will be our subject here.

Unless the detective finds the clues, judge or jury has nothing to consider. **Unless exploratory data analysis uncovers indications, usually quantitative ones, there is likely to be nothing for confirmatory data analysis to consider.**

Experiments and certain planned inquiries provide some exceptions and partial exceptions to this rule. They do this because one line of data analysis was planned as part of the experiment or inquiry. **Even here, however, restricting one's self to the planned analysis--failing to accompany it with exploration--loses sight of the most interesting results too frequently to be comfortable.**

As all detective stories remind us, many of the circumstances surrounding a crime are accidental or misleading. Equally, many of the indications to be discerned in bodies of data are accidental or misleading. To accept all appearances as conclusive would be destructively foolish, either in crime detection or in data analysis. **To fail to collect all appearances because some--or even most--are only accidents would, however, be gross misfeasance deserving (and often receiving) appropriate punishment.**

Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone--as the first step.

We will be exploring numbers. We need to handle them easily and look at them effectively. Techniques for handling and looking--whether graphical, arithmetic, or intermediate--will be important. The simpler we can make these techniques, the better--so long as they work, and work well. When details make an important difference, they deserve--and will get--emphasis.

review questions

What is exploratory data analysis? How is it related to confirmatory data analysis? How is preplanned analysis related to exploratory data analysis? Should we look only at appearances we are sure are correct?