13

EXPLORATORY DATA ANALYSIS: AN INTRODUCTION TO SELECTED METHODS

Samuel Leinhardt
CARNEGIE-MELLON UNIVERSITY

Stanley S. Wasserman
UNIVERSITY OF MINNESOTA

The effective analysis of quantitative empirical data typically entails a two-stage process in which data are first surveyed and explored using "quick and dirty" methods, and then,

This report is part of a continuing collaboration in which the authors share equal responsibility. As a consequence, their names appear in alphabetical order.

Partial support for this study was provided by a grant from the U.S.

once patterns have been made evident, models are evaluated and hypotheses tested using the procedures of classic inferential statistics. Most experienced data analysts employ such a compound procedure even when their objective is to investigate either theoretically well-specified functional dependencies among variables or the results of carefully planned experiments. Nonetheless, the first phase of this process, the exploratory phase, is rarely if ever reported. There are two reasons for this lack of reportorial attention to the preliminary exploration of data. First, students of data analysis sometimes are taught that extensive exploration of data prior to the application of confirmatory tools is somehow suspect, representing a form of methodological cheating. Second, many of the exploratory procedures investigators employ are of an ad hoc nature and are seemingly devoid of an underlying theoretical rationale. Not willing to have perfectly reasonable findings dismissed because the data were used to inform their own analysis or because procedures were used that are not readily justified, many analysts engage in extensive exploration of data but do not report these activities.

John W. Tukey, of Princeton University and Bell Laboratories, has formulated a systematic approach to exploratory data analysis (EDA) that promises to bring this phase of data analysis out of the closet and into the limelight. Tukey (1977) has developed a large and diverse array of intuitively sound and theoretically supportable tools for exploring data prior to the application of confirmatory methods. These tools help investigators organize data efficiently, construct compelling graphic displays, examine traditional distributional assumptions, and explore the structure of functional dependencies—all without recourse to the probabilistic assumptions basic to traditional statistical methods. Moreover, they provide a means for remedying such common data problems as stray data values, asymmetry, nonlinearity, and location-scale interaction, problems which, if left unresolved, ser-

Department of Housing and Urban Development. Paul W. Holland played an essential role in suggesting some of the ideas developed in this report. We are also grateful to Gaea Leinhardt, David C. Hoaglin, Joseph B. Kadane, and the staff of the Quantitative Methods for Public Management Curriculum Development Project. An anonymous referee also provided useful suggestions.

iously impair the effectiveness of most confirmatory procedures. Because Tukey's methods possess a high degree of resistance (results are little affected by large deviations in a small subset of data values) and some robustness (results are little affected by changes in underlying distributions) they are well suited for preliminary analyses—especially when, as in most empirical sociological studies, substantive theory provides little more than a hint of the functional dependence to be modeled. Exploratory data analysis, as Tukey envisages it, is essentially a process through which data are examined to gain insight into how variables should be expressed, how the functional structure should be modeled, and what analytic methods should be used to test and evaluate the model. The vast array of tools he has created facilitates this process, renders it consistent, and thus possesses the potential of significantly improving the analysis of quantitative empirical data.

During the last 20 years, work in this area has appeared in technical journals and books far removed from the literature normally scanned by sociologists. Although a preliminary version of Tukey's text was circulated in the early 1970s, its unique organization, peculiarly personal style, large size, and lack of theoretical depth caused his methods to remain arcane and little known to nonstatisticians throughout this period. His recently published text (Tukey, 1977), although a reduction in scope from the preliminary edition, has done little to alter this situation. The treatise still possesses the unique style of the preliminary edition and, consequently, is a difficult book for professionals, let alone beginning students.

This chapter is our attempt to improve upon this situation by introducing sociologists to EDA techniques and providing rationales for their application. In so doing, we wish to demonstrate the analytic utility of these specific tools and argue for the general approach of EDA—namely, using the data to inform the choice and process of an analysis. Moreover, we wish to make a case for the use of EDA as a pedagogic tool for introducing students to quantitative data analysis and the construction of mathematical models. Although Tukey's writing style may make reading his text arduous, it is our firm belief, one which has been justified in the classroom, that teaching EDA before classic inferential statistics is an effective pedagogic strategy. Alternative texts by

other authors may ultimately ease the student's burden. (See the discussion section on page 361 of this chapter.)

This chapter is organized as follows. We introduce a data set, the "nations data," which we shall use in one form or another throughout the discussion. Each section presents an EDA procedure for exploring increasingly more complicated data structures. These methods are then applied illustratively. As the structure of the data increases in complexity, more complicated tools are needed and more sophisticated models are developed.

David (1977), Wainer (1977), Kadane (1978), and Beniger (1978) have presented critical reviews that evaluate Tukey's text from several perspectives (see also Welsch, 1978). Such is not our purpose. Instead, we have chosen to select a variety of methods Tukey has pioneered and illustrate their application in order to provide an effective introduction to Tukey's approach. Readers interested in using Tukey's book for research or academic purposes will find the aforementioned reviews useful.

EDA FOR SOCIOLOGISTS

The techniques of EDA possess features that render them quite useful in many of the analytic situations faced by empirical sociologists. Most empirical sociological research, if not completely atheoretical, is guided at best by vague notions concerning which variables should be included in an explanatory model. The absence of compelling theory, particularly evident in applied sociology, program evaluation, and policy studies, is compounded by the fact that often the empirical scientist or policy analyst must resort to a study of someone else's data. Such secondary data analyses usually require the analyst to use measures that rarely correspond to theoretical variables. Moreover, such analyses often have relevant observations missing, involve measurements made among arbitrarily chosen scales, and have data organized in an analytically arbitrary manner.

Exploratory data analysis can be of great assistance when such secondary analyses are undertaken. It provides methods for organizing and summarizing data, for detecting data characteristics that invalidate the usual statistical assumptions, for constructing reasonable models in the absence of directing theory, and for estimating parameters even in the presence of deviant or

missing values. Thus it provides the empirical sociologist with a new way to attack data, one that is particularly effective and relevant.

While EDA procedures possess high utility in analyzing secondary or poor-quality data or when theory does not provide insight into the exact specification of a model, they can still play an important role when more favorable conditions prevail. When, for example, a carefully constructed experiment has supposedly predetermined the nature of the analysis, or when theory guides the initial choice of the functional form of an equation for modeling the data, the use of EDA procedures often yields more effective functional forms or scale adjustments that make for a more productive analysis and, consequently, greater understanding. Often, too, unforeseen results emerge that might otherwise have been entirely overlooked.

Exploratory data analysis has another potentially far-reaching role to play. Because of its emphasis on graphics, lack of distributional assumptions, and use of iterative fitting procedures, EDA represents an unexcelled vehicle for introducing sociology students to quantitative data analysis and mathematical modeling. As an introduction to traditional statistical ideas, EDA provides students with a set of easily understood tools whose quality of "face validity" enables them to amass a solid base of analytic experience before encountering the abstract notions of traditional statistical inference. Moreover, a solid grounding in EDA often provides students with a healthy attitude toward data; they learn that data are not sacrosanct and may require adjustments before effective analysis can commence. By discovering that the data can provide the analyst with information about how they should be analyzed, students learn to eschew blind mechanical application of analytic procedures and develop, instead, an inquisitive, analytic approach that promotes the creative exploitation of empirical data and helps them understand model-building activities in general (Leinhardt and Wasserman, 1978; in press).

TOOLS FOR SINGLE BATCHES

Among his many other contributions, Tukey has introduced a specialized terminology that some will rue and others will find intuitively appealing. Much like the proverbial scorecard at a

baseball game, however, one cannot read about EDA without a functional knowledge of Tukey's terms. We introduce some of these special terms as we proceed, but it may be useful to consider briefly why they are needed.

With many EDA tools, such as the *stem-and-leaf display* described later on, no classic equivalent exists, and a name for the display is obviously required. In this case, as in all others, Tukey has attempted to create a name that is a simple description of the procedure or a mnemonic for recalling a quality of either the procedure or some part of the procedure. With some terms, such as the *hinge,* an approximate quartile, the reference is obscure, and some may question its mnemonic utility. With other terms, such as *fence* and *comparison value,* the names are readily accepted once the roles of the procedures are understood. In several cases, there is an existing classic term referring to a concept quite similar to an EDA term. The term *hinge* is an example. It is a sample statistic that is approximately equal to a quartile, an estimate of the 25th and 75th percentage points of an underlying probability function. However, it is important to emphasize that, while similar, the classic and EDA calculations may yield different results in some circumstances. Thus, to assure consistency and comprehension, separate terms should be used. Indeed some EDA authors, who are dissatisfied with Tukey's choice of words, have adopted alternative terminology, such as the terms *quarter* or *fourth* for *hinge* (Hoaglin, in press). Sometimes, as with the term *quarter,* the change increases consistency. Quarters lead naturally to eighths, sixteenths, and so on, all of which are sample order statistics computed in a manner consistent with the computation of a quarter. While a consistent nomenclature is desirable, none has evolved as of this writing; therefore in what follows, we utilize Tukey's terminology and give alternative terms that have come to our attention.

Stem-and-Leaf Displays

The first concept we introduce in a quasi-formal manner is a *batch*—a set of similar values collected in some consistent fashion. This notion of a batch of data is fundamental and is kept purposefully vague. It is the EDA equivalent of the classic definition of a data set as a collection of empirical realizations of a ran-

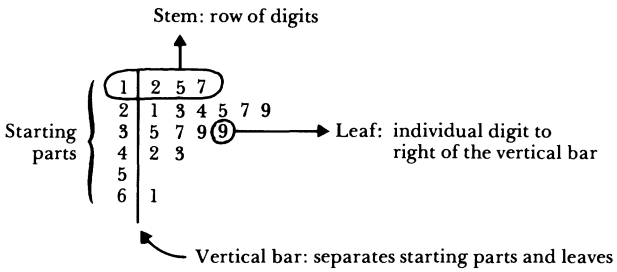
dom variable. Indeed, one can consider a batch as simply a data set consisting of a single variable. The vagueness of the term *batch* is purposeful in that it does not depend on the definition of a variable, the meaning of *realization,* or even the notion, implicit in the more traditional view, of a sample, random or otherwise. Batches are collections of values that data analysts find useful to study. Batches may occur singly, as simple sets of observations on a single variable, or as multiple ordered batches, or related sets of observations on several different variables.

Single batches are the least complicated data structure. The EDA procedures for single batches consist of an initial sequence of activities in which data are organized to facilitate analysis, condensed, and then summarized prior to the construction of a model. Systematic organization and condensation are necessary to promote operational efficiency in any investigation into the behavior of data values. The stem-and-leaf display (Tukey, 1972) is a basic organizational device of EDA. Like a simple sorting of data values, the display organizes the values in numerical order. And, like a histogram, the display facilitates study of the shape, spread, and distributional characteristics of the empirical values. Unlike both a sort and a histogram, the stem-and-leaf display is quickly and easily constructed by hand. In addition, it retains information on individual data values and provides a convenient medium for display or storage.

A stem-and-leaf display is composed of a column of digits called starting parts (or stem values) that is separated by a vertical line from rows of digits. Each row is a stem; the individual digits on it are leaves. Construction of the display involves decomposing each datum into a stem value and a leaf. For example, assume that the ordered data values[1] 12, 15, 17, 21, 23, 24, 25, 27, 29, 35, 37, 39, 39, 42, 43, 61 constitute a batch. One possible decomposition is to use each 10's digit as a stem value or starting part and the unit's digit as a leaf. Thus 12 would decompose into 1 for the starting part and 2 for the leaf. The completed display of these data values would appear as in Figure 1. Note that those values with the same 10's digit share the same starting part. To recon-

[1]Values have been ordered for illustrative purposes. The procedure is identical with unordered values.

Figure 1. Simple stem-and-leaf display.



NOTE: Unit = 10°.

struct the actual data values from the display, one simply juxtaposes a leaf to the immediate right of its respective starting part and multiplies the juxtaposed digits by the indicated scale unit (upper left-hand corner of Figure 1). For example 2 positioned to the right of the 1 starting part yields $12 \times 10^0 = 12$ for the first value of the display. Proceeding similarly, one can reconstruct each of the 16 original values in this data set. Note that other decompositions of the data values are possible and, in certain instances, preferable (see Tukey, 1977).

Several features of the display should be noted. First, the display is economical; 22 single digits have been used to represent 16 two-digit numbers. Much greater economies occur when more extensive data sets are analyzed. A data value with many digits can be reduced to only a few digits in the display. The individual values are retained or, at least, can be readily reconstituted. The values are easily ordered (although to do so from raw data requires constructing two stem-and-leaf displays), and the leaves on each stem can be counted, thus facilitating the location of order statistics such as the quartiles and the median. The display also focuses attention on shape. Since each leaf occupies a set amount of space in the display, the length of a line is proportional to the number of data values. Besides general impressions regarding shape, spread (or variation), and center, the display clearly indicates where clustering occurs.

Like a histogram, there are many varieties of stem-and-leaf

displays that differ primarily in the number of stems associated with each starting part. Increasing the number of stems spreads out the display; decreasing them makes the display more compact. As a general rule for batches with N values: Displays should have no more than $10(\log_{10}N)$ stems or lines (Hoaglin and Wasserman, 1975).

Table 1 presents data drawn from the editorial section of *The New York Times* (Crittenden, 1975). The data are estimates of infant mortality rates, the number of deaths per 1000 live births, for 105 nations of the world, part of the "nations" data set. The data are organized in an analytically arbitrary manner based on the alphabetical ordering of the nations' names.

Figure 2 contains a stem-and-leaf display of the data in Table 1. The unit chosen for the leaves is 10^0. The data values have been truncated or cut (not rounded) before being decomposed for placement in the display. A version of the display with two-digit leaves could be constructed if retention of each given data value was deemed essential. To reconstruct the cut values of the display,

Figure 2. Stem-and-leaf display of infant mortality rates by nation.

1	0	9	
19	1	0 0 1 1 1 2 2 3 3 5 6 6 6 7 7 7 7 9	
31	2	0 0 1 1 2 3 5 6 6 7 7 8	
35	3	2 4 8 9	
41	4	0 3 4 5 8 8	
49	5	0 1 4 5 8 8 8 9	
(7)	6	0 0 0 2 3 5 7	
45	7	6 8 8 9	
41	8	0 4 6	
38	9		
38	10	0 0 2 4 9	
33	11	4	
32	12	0 4 4 5 7 9	
26	13	2 7 8 9	
22	14	8 9	
20	15	0 3 8 9	
16	16	0 0 2	
13	17	0 0	
11	18	0 0 7	
8	19	0	
7	20	0 0	
5	21	6	
4	HI	259, 300, 400, 650	Zambia, Libya, Afghanistan, Saudi Arabia

NOTE: Unit = 10^0, values cut.

TABLE 1
Infant Mortality Rates for 105 Nations (circa 1970)

Nation	Rate	Nation	Rate	Nation	Rate
1. Afghanistan	400.0	37. India	60.6	73. Portugal	44.8
2. Algeria	86.3	38. Indonesia	125.0	74. Rwanda	132.9
3. Argentina	59.6	39. Iran	NA	75. Sierra Leone	170.0
4. Australia	16.7	40. Iraq	28.1	76. Singapore	20.4
5. Austria	23.7	41. Ireland	17.8	77. Somalia	158.0
6. Bangladesh	124.3	42. Israel	22.1	78. South Africa	71.5
7. Belgium	17.0	43. Italy	25.7	79. Saudi Arabia	650.0
8. Bolivia	60.4	44. Ivory Coast	138.0	80. South Korea	58.0
9. Brazil	170.0	45. Jamaica	25.2	81. South Yemen	80.0
10. Britain	17.5	46. Japan	11.7	82. Spain	15.1
11. Burma	200.0	47. Jordan	21.3	83. Sri Lanka	45.1
12. Burundi	150.0	48. Kenya	55.0	84. Sudan	129.4
13. Cambodia	100.0	49. Laos	NA	85. Sweden	9.6
14. Cameroon	137.0	50. Lebanon	13.5	86. Switzerland	12.8
15. Canada	16.8	51. Liberia	159.2	87. Syria	21.7
16. Central African Republic	190.0	52. Libya	300.0	88. Taiwan	19.1
17. Chad	160.0	53. Madagascar	102.0	89. Tanzania	162.5

18. Chile	78.0	54. Malawi	148.3	90. Thailand	27.0
19. Colombia	62.8	55. Malaysia	32.0	91. Togo	127.0
20. Congo	180.0	56. Mali	120.0	92. Trinidad	26.2
21. Costa Rica	54.4	57. Mauritania	187.0	93. Tunisia	76.3
22. Dahomey	109.6	58. Mexico	60.9	94. Turkey	153.0
23. Denmark	13.5	59. Morocco	149.0	95. Uganda	160.0
24. Dominican Republic	48.8	60. Nepal	NA	96. United States	17.6
25. Ecuador	78.5	61. Netherlands	11.6	97. Upper Volta	180.0
26. Egypt	114.0	62. New Zealand	16.2	98. Uruguay	40.4
27. El Salvador	58.2	63. Nicaragua	45.0	99. Venezuela	51.7
28. Ethiopia	84.2	64. Niger	200.0	100. Vietnam	100.0
29. Finland	10.1	65. Nigeria	58.0	101. West Germany	20.4
30. France	12.9	66. Norway	11.3	102. Yemen	50.0
31. Ghana	63.7	67. Pakistan	124.3	103. Yugoslavia	43.3
32. Greece	27.8	68. Panama	34.1	104. Zaire	104.0
33. Guatemala	79.1	69. Papua	10.2	105. Zambia	259.0
34. Guinea	216.0	70. Paraguay	38.6		
35. Haiti	NA	71. Peru	65.1		
36. Honduras	39.3	72. Philippines	67.9		

NOTE: NA = missing value.

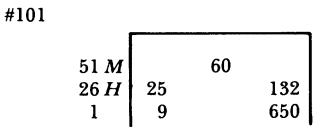
we proceed as before, taking a leaf—for example, the 9 in the first stem juxtaposed with the starting part, 0—and multiplying by the unit 10^0 to form an infant mortality rate of 9. Referring to the data reveals that this data value is exactly 9.6: Sweden. Other examples of reconstruction include 216 for the largest rate appearing in the display and 60 for the median value. Note the column of cumulative counts to the left of the display. The counts or *depths* are recorded in such a manner that they increase from either extreme toward the middle or median value. The stem containing the median has an uncumulated count in parentheses so that the median data value is readily obtained. Note also that to avoid displaying many starting parts devoid of leaves before reaching the large observed values, the display has a separate row labeled "HI" for the large outlying data values. A similar procedure would be used if there were outlying data values at the low end. The asymmetry of these values, trailing out toward extremely high infant mortality rates, is clearly apparent, as is the clustering of values between 10 and 30.

Five-Number Summaries and Letter-Value Displays

While the stem-and-leaf display is a readily constructed and easily appreciated device for storing and displaying data, it usually contains too much information for further analysis. Condensing this information into a standard set of summary statistics that provide information on typical value (or *location*) and spread (or *dispersion*) of the batch is often essential. Expanding this set to include indicators of behavior in the tails of the empirical distribution is often useful for distinguishing between various distributional forms and for identifying potentially deviant or suspect values.

Numerous statistics could serve these purposes. The EDA procedures, since they are designed to handle non-Gaussian data, typically involve the use of order statistics (such as the median) that are relatively resistant. Resistance is a quality possessed by statistics that are little affected by highly deviant values. (An example of a nonresistant statistic is the mean; see Mosteller and Tukey, 1977.) Tukey calls the order statistics *letter values,* which he organizes into a *letter-value display,* since letters are used as mnemonics for the statistics.

Figure 3. Simple letter-value display of infant mortality data from the nations data set.



The depth of the median is $(N + 1)/2$, where N is the number of observations in the batch. When N is odd, this formula yields an integer, and the median is located by counting in from either end of the sorted batch to the depth indicated by the formula. When N is even, the result is a fractional depth halfway between the two middle values. In this instance, the median is computed by taking the arithmetic average of these two values, and its value does not correspond to any data value in the batch. This general definition of a middle value can be extended to the two halves, four quarters, and so forth of the batch. For example, since the median cuts a batch in half, one could find the medians of these two halves, or approximate quartiles. These quartile statistics would be located at depth

$$\frac{1 + \lfloor \text{depth of median} \rfloor}{2}$$

(The symbol $\lfloor\ \rfloor$ represents the mathematical *floor function,* which merely truncates positive reals to the next smaller integer.) Tukey calls these empirical quartiles *hinges,* located by counting in the same depth from either extreme. One could continue this process to locate the eighths at depth = $(1 + \lfloor \text{depth of hinge} \rfloor)/2$, the sixteenths at depth = $(1 + [\text{depth of eighth}]/2$, and so on.[2]

Letter-value displays are used to organize these summary statistics. Figure 3 presents a letter-value display containing the statistics from a five-number summary, the median M, the hinges H, and the extremes. The basic objective is to provide the overall count of values (101 instead of 105 because 4 NAs are excluded) and depths outside the three-sided box and the corresponding

[2]We refer to "approximate" or empirical quartiles because the truncation rule applied here may yield statistics that differ from those found by using more mathematical definitions of order statistics.

Figure 4. Letter-value display for infant mortality data.

#101

			60	
51	*M*			
26	*H*	25		132
13.5	*E*	16		166
7	*D*	12		200
4	*C*	11		259
2.5	*B*	10		350
1		9		650

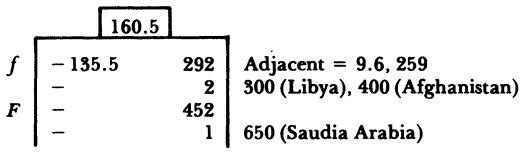
data values inside. Figure 3 is a simple letter-value display for the infant mortality data. Note that the extremes are always at depth 1. A more complete letter-value display, as in Figure 4, includes additional order statistics such as the eighths (E), sixteenths (D),[3] and thirty-secondths (C), which emphasize the behavior of the tails of the batch.

One of the principal uses of these displays and statistics is to compute other characteristics of the data. For example, the midspread or hinge spread (Q spread for Hoaglin because it is the approximate interquartile range) is the distance between the hinges (that is, dH = upper hinge − lower hinge) and is a useful measure of spread. In a Gaussian batch, $\frac{3}{4}dH \approx$ standard deviation (precisely, $dH = 1.349\sigma$). In a non-Gaussian, asymmetric batch, the standard deviation is of questionable utility in describing spread or variation; however, the midspread always indicates the length of the interval within which the central 50 percent of the data values fall. Spreads can be included in a letter-value display by placing them in a column to the immediate right of the box.

The relationship between the hinge spread and the standard deviation can be used to motivate the use of the hinge spread as an indicator of deviant values. Tukey defines a *step* as equal to $1.5dH$ and suggests that any data value in a batch which lies one step beyond either hinge should be viewed as stray. Values two steps or more out are especially deviant. Tukey constructs two sets of hypothetical *fences* at these locations and uses these bound-

[3]Instead of trying to maintain first-letter mnemonics, Tukey switches to a sequence based on letters of the alphabet for the less commonly used order statistics from sixteenths, D, through 128ths, A, moving in reverse alphabetical order.

Figure 5. Fenced letter-value display addition for infant mortality data.



ing markers to draw attention to potentially questionable values. Those fences—one step out on either side—are called the *inner fences;* those two steps out are the *outer fences*. The single values nearest each inner fence on the inside (that is, the side toward the median) are termed *adjacent* because they are adjacent to the fence. Computed numerical values of the fences of a batch can be displayed in a *fenced letter-value display,* constructed by adding the component indicated in Figure 5 to that of Figure 3 or Figure 4.

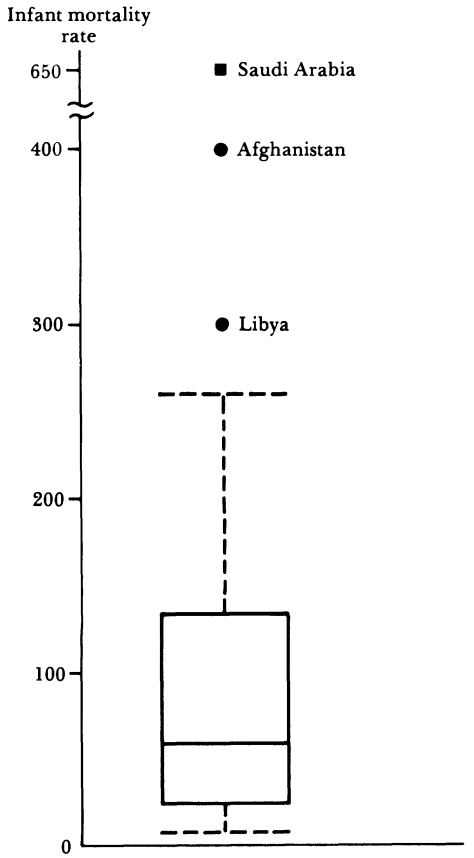
The enclosed value at the top of this display, 160.5, is the value of a step in the infant mortality data batch. The lowercase *f* indicates the inner fence; the uppercase *F* indicates the outer fence. Computed values for the inner and outer fences are displayed on the respective rows within the box. Immediately below these values are the number of actual data values between the inner and outer fences and outside the outer fences. For example, the upper inner and outer fences are calculated to be 292 and 452, respectively, for the infant mortality data. Two data values fall between these fences, and one data value lies beyond the upper outer fence. In this batch there are no data values beyond the lower fence. Any values falling between the two fences are called *outside;* those beyond the outer fence are *far outside*. These outside and far outside values are given to the right of the display, along with the nations they refer to. Were this a substantive analysis, the statistics would be interpreted to imply that these values are quite different from the bulk of other observations. If they are truly outliers—that is, if they remain deviant even after a transformation—it might be wise to exclude them from further analysis and study them separately so that they would not exert undue influence (as would be the case if the data analysis involved use of a nonresistant technique, such as analysis of variance or ordinary least-squares regression).

Schematic Plots

Five-number summaries constitute the initial stage of an exploratory analytic condensation. We shall shortly expand upon their role. Often it is convenient to have a graphic representation of the information they contain. The display that accomplishes this is called a *schematic plot*. As its name implies, it graphically schematizes the summary statistics of a batch. Unlike the stem-and-leaf display, however, it is a condensation of the data and thus involves a loss of information. In particular, clustering and multimodality of data values are not always visible in schematic plots.

Figure 6 is a schematic plot of the infant mortality data by

Figure 6. Schematic plot of infant mortality rates.



nation. The vertical scale is the infant mortality rate. The box in the display encompasses the central 50 percent of the values and extends from the lower hinge to the upper hinge. The horizontal line across the box indicates the median.[4] Vertical dashed lines extend from each hinge to the adjacent values, which are indicated by horizontal dashed lines. Outside values are individually noted by solid circles, and the far outside value is noted by a solid square. Obviously the width of the display is arbitrary, and the orientation of the display, vertical or horizontal, is simply a matter of taste or convenience. When the procedure is computerized, horizontal rather than vertical displays permit ready comparisons across schematics of different batches.

Transformations for Symmetry

The approximate equality between $\frac{3}{4}dH$ and the standard deviation in a symmetric, well-behaved (Gaussian) batch motivates the use of the hinge spread in partitioning a batch in regions of acceptable and suspect data values. In a well-behaved batch, for example, $1.5dH$ or one step is approximately equal to 2 standard deviations. Moreover, the hinge (to which one adds or subtracts a step to find the inner fence) is $\frac{1}{2}dH$ or $\frac{2}{3}$ standard deviation from the median, which, in this symmetric case, is the same as the mean. Thus the inner fence is approximately $2\frac{2}{3}$ standard deviations from the mean and the outer fence approximately $4\frac{2}{3}$ standard deviations away. With a standard normal distribution, the inner fences would bound approximately 99 percent of the data and the outer fences would bound very nearly 100 percent. Clearly values lying between the fences should be rare, and values lying beyond the outer fences should be exceedingly rare.

When a batch is symmetric, the use of classic summary statistics such as the mean and standard deviation may be appropriate, and when the batch is well-behaved (that is, Gaussian) these statistics are sufficient (in a statistical sense) to characterize it. Thus if an asymmetric, ill-behaved batch can be made well behaved through some simple and meaningful algebraic operation, the gain in analytic summarization may be substantial.

[4]Another line, distinguished from that indicating the median, might be provided to indicate the mean. However, in certain batches the mean will not fall within the box.

In EDA, power transformations are used to attain this goal. Simple power transformations retain the original sequence of data values up to a change in sign so that order statistics deduced from the original untransformed data mark the same positions through all transformations and vice versa. Tukey develops the logic of using transformations to obtain symmetry by examining the impact of various power transformations on empirical distributions.

One mathematical approach to the use of symmetrizing transformation that we have adopted follows. Given a variable Y, restricted to the positive real numbers, consider the parametric family of transformations defined as

$$Y(p) = \begin{cases} (Y^p - 1)/p & \text{if } p \neq 0 \\ \log Y & \text{if } p = 0 \end{cases} \tag{1}$$

which is a continuous function of p at $p = 0$.[5] A batch of data is considered symmetric if the distances from the hinges to the median are equal. (Strictly speaking, equality of differences between the eighths and the median, sixteenths and the median, and so forth is also required for symmetry. See Hinkley, 1975, for a maximum-likelihood procedure to combine these various criteria.) Thus we examine

$$M - LH = UH - M \tag{3}$$

where UH and LH are the upper and lower hinges (or quarters) and M is the median of the batch. We use the simple power transformation Y^p when $p \neq 0$, and $\log Y$ when $p = 0$, instead of $Y(p)$ for mathematical ease. The symmetrizing transformation is

[5]One justification for using $\log Y$ when $p = 0$ follows. Note that

$$Y(p) = (1/p)(Y^p - 1) = \int_1^Y x^{p-1}\,dx$$

and that

$$Y(0) = \int_1^Y (1/x)\,dx = \lim_{n \to \infty} \sqrt[n]{Y - 1} = \log Y \tag{2}$$

by definition. Thus the assumption that $\log Y$ is the zeroth power is true by a simple result from calculus. It may also be seen geometrically by examining Tukey's *ladder of powers*.

a specific power of Y such that

$$Y_M^p - Y_{LH}^p = Y_{UH}^p - Y_M^p \tag{4}$$

where Y_M indicates the median, Y_{LH} and Y_{UH} the hinges, of a sample from the random variable Y. Equivalently,

$$Y_M^p = \tfrac{1}{2}(Y_{LH}^p + Y_{UH}^p) \tag{5}$$

We now give a rule for finding p that utilizes a Taylor series expansion of Y^p.[6] (The nontechnically oriented reader may skip the mathematical derivations. An approximation formula for transformation is given in Equation 15.)

If $\tilde{Y}$ is in the neighborhood of Y, then [7]

$$f(Y) = f(\tilde{Y}) = f'(\tilde{Y})(Y - \tilde{Y}) + o[(Y - \tilde{Y})^2] \tag{6}$$

so that if $f(Y) = Y^p$, we have

$$Y^p \approx \tilde{Y}^p + p\tilde{Y}^{(p-1)}(Y - \tilde{Y}) \tag{7}$$

Therefore

$$\begin{aligned} Y_M^p - Y_{LH}^p &= Y_L^p + pY_L^{p-1}(Y_M - Y_L) \\ &\quad - [Y_L^p + pY_L^{p-1}(Y_{LH} - Y_M)] = pY_L^{p-1}(Y_M - Y_{LH}) \end{aligned} \tag{8}$$

where Y_L lies between Y_M and Y_{LH}.

Similarly,

$$Y_{UH}^p - Y_M^p = pY_H^{p-1}(Y_{UH} - Y_M) \tag{9}$$

where Y_H lies between Y_M and Y_{UH}.

Our criterion (4) becomes

$$pY_L^{p-1}(Y_M - Y_{LH}) = pY_H^{p-1}(Y_{UH} - Y_M) \tag{10}$$

or

$$(Y_L/Y_H)^{p-1} = (Y_{UH} - Y_M)/(Y_M - Y_{LH}) \tag{11}$$

Thus the approximate power for transformation is

$$p_0 = 1 + \log\left(\frac{Y_{UH} - Y_M}{Y_M - Y_{LH}}\right) \Big/ \log\left(\frac{Y_L}{Y_H}\right) \tag{12}$$

[6] From a suggestion by P. W. Holland.

[7] The notation $o[(Y - \tilde{Y})^2]$ stands for additional terms in the expansion that become negligible for large $(Y - \tilde{Y})^2$ and hence can be ignored. See Bishop and others (1975, chap. 14) for a discussion of o.

As a further simplification, we can take

$$Y_L = \tfrac{1}{2}(Y_{LH} + Y_M) \tag{13}$$

and

$$Y_H = \tfrac{1}{2}(Y_M + Y_{UH}) \tag{14}$$

The operation then becomes: Compute Y^{p_0}, where

$$p_0 = 1 + \log\left(\frac{Y_{UH} - Y_M}{Y_M - Y_{LH}}\right) \Big/ \log\left(\frac{Y_{LH} + Y_M}{Y_M + Y_{UH}}\right) \tag{15}$$

as an approximation to the correct power transformation and examine the data to determine if symmetry is achieved. Since the solution is approximate, p_0 should be used as a guide with slightly higher or lower values of p_0 tried if circumstances direct. Hinkley (1977) gives an even simpler criterion based on a standardized difference between the sample mean and median to find p_0.

Tukey has not yet presented a good rule of thumb for finding reasonable p's. He prefers to use power transformations drawn from his *ladder of powers*—essentially a small set of mostly integer values of p ranging from the negative reciprocal cube (negative to maintain the original order of the batch values) through the cube, but including the square root and negative reciprocal square root and the log when p equals zero. His ex-

TABLE 2
Transformation Summaries for Infant Mortality Rates

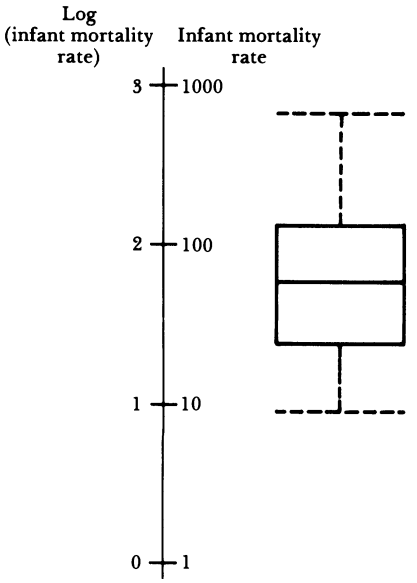
	Raw Data $p = 1$	Square Roots $p = 0.5$	Logarithms $p \approx 0.0$	Negative Reciprocal Roots $p = -0.5$
Min	9	3.0	0.95	−0.33
E	16	4.0	1.20	−0.25
H	25	5.0	1.40	−0.20
M	60	7.75	1.78	−0.13
H	132	11.49	2.12	−0.09
E	166	12.88	2.22	−0.08
Max	650	25.50	2.81	−0.04
M	60	7.75	1.78	−0.13
mid *H*	78.50	8.25	1.76	−0.14
mid *E*	91.00	8.44	1.71	−0.17
mid extreme	329.50	14.25	1.88	−0.19
Trend	(up)	(slightly up)	(flat)	(slightly down)

ploratory procedure for selecting p_0 takes advantage of the fact that these power transformations preserve order. Thus one simply constructs a five-number summary of the data, tries various transformations, and examines trends in the differences between the transformed median and the midsummaries (estimates of the center constructed from the average of paired transformed order statistics), the midhinge (the average of the transformed hinges), the midextreme, and so on. The absence of a trend (that is, when the median, midhinge, midextreme, and so on are roughly equal) indicates that the batch is reasonably symmetric, and the transformation that led to this situation is chosen as the symmetrizing transformation. Table 2 illustrates the procedure with the infant mortality data.

Tukey's procedure directs the analyst to try powers on the ladder of powers, moving in a direction opposite to the apparent trend in the transformation summary. Thus in Table 2 the trend in the raw data ($p = 1$) is up as one goes from the median to estimates of the center based on averages of more extreme order statistics. The first try here is the square root ($p = \frac{1}{2}$). The summary is improved but still has an upward trend. When logarithms ($p \approx 0$) are used, there is no trend; negative reciprocal square roots yield a downward trend and thus are too far down the ladder of powers. The choice is the logarithm. Equation (15) yields $p_0 = 0.1$ directly, quite close to the choice of $p_0 \approx 0$ or the logarithm. Figure 7 presents a schematic plot of the infant mortality data transformed by taking logarithms. Notice that there are no longer outliers present, a finding that suggests the data should be thought of on a logarithmic scale.

Transformations play an absolutely fundamental role in EDA. One possible view is to consider them to be scale adjustments under the assumption that the scale in which the data were originally measured may not necessarily be the scale in which the underlying variable should be expressed for a parsimonious summarization. Sociology does not yet possess sufficient theoretical development to permit the application of procedures (such as dimensional analysis in physics) by which the proper scale can be determined *a priori*. The EDA procedures operate so as to draw information on scale transformations from the data. Other uses of transformations, to be detailed below, include the achievement

Figure 7. Schematic plot of logarithms of infant mortality data by nation.



of constancy in variation, additivity, and linearity. All these operations act, in the long run, to increase the utility and efficiency of traditional statistical methods.

Scale-adjusting transformations are not new to sociologists. Standardization, the construction of standard normal deviates, is a well-established procedure. The objective of this operation is to reexpress the raw data values of a normally distributed variable in a scale in which the unit is the standard deviation. Such a transformation provides a change in perspective, a new way of thinking about the data that facilitates their analysis. The objective in finding a symmetrizing transformation is similar. We seek a simple algebraic operation that alters the scale of the data, making it more comprehensible and more amenable to well-known analytic procedures. If we conceive of our analytic objective, even in the case of observations on a single variable, as the fitting of a model to a data set, then we can think about the center of the batch as our "fit" and deviations from it as residuals (as in the general equation: data = fit + residual). If the data are symmetric and well behaved, these residuals will be symmetrically

distributed around zero with characteristics similar to those we wish to see when analytic procedures require distributional assumptions about the probabilistic nature of error terms.

A compelling rationale can be constructed for seeking symmetrizing transformations for single batches. In more complicated data structures, other objectives, to be detailed below, take clear priority. However, the process of finding a symmetrizing transformation is very important pedagogically. Usually it represents the student's first exposure to scale adjustments and thus is a simple introduction to the fact that the values used in a final analysis may *not* be the original measured values but some function of them. It also familiarizes students with "foreign" or unnatural units such as the square root of income or the logarithm of population. By exposing sociology students to the use of such mathematical entities as logs and exponents, symmetrizing transformations help them discover that data analysis involves finding relationships among variables that can be expressed mathematically. With this goal in mind, sociological data are no different conceptually from empirical observations in any scientific discipline and may require algebraic manipulation regardless of whether they describe income, mortality rates, sizes of ethnic groups, or rates of interregional migration.

TOOLS FOR MULTIPLE UNORDERED BATCHES

Suppose that one has a collection of single batches that are related in some qualitative way which does not permit an ordering of the batches. We call this data structure a collection of multiple unordered batches. Examples include the 1970 populations of census tracts in Pittsburgh and Omaha, achievement test scores for first-year graduate students categorized by undergraduate major course of study, and life expectancies for nations of the world classified by region.

The graphic EDA procedures useful for the initial analysis of multiple unordered batches include stem-and-leaf displays and schematic plots drawn in parallel, either side by side or one above the other. To improve the effectiveness of such displays when the spreads in the individual batches vary greatly, Tukey has developed a procedure for determining a transformation that stabilizes

TABLE 3

Infant Mortality Rates for Nations of the World Classified by Region

The Americas		Africa		Europe		Asia-Oceania	
Canada	16.8	South Africa	71.5	Austria	23.7	Australia	16.7
United States	17.6	Algeria	86.3	Belgium	17.0	Japan	11.7
Ecuador	78.5	Libya	300.0	Denmark	13.5	New Zealand	16.2
Venezuela	51.7	Nigeria	58.0	Finland	10.1	Indonesia	125.0
Argentina	59.6	Tunisia	76.3	France	12.9	Iran	NA
Brazil	170.0	Zambia	259.0	West Germany	20.4	Iraq	28.1
Chile	78.0	Cameroon	137.0	Ireland	17.8	Saudi Arabia	650.0
Colombia	62.8	Congo	180.0	Italy	25.7	Israel	22.1
Costa Rica	54.4	Egypt	114.0	Netherlands	11.6	Lebanon	13.6
Dominican Republic	48.8	Ghana	63.7	Norway	11.3	Malaysia	32.0
Guatemala	79.1	Ivory Coast	138.0	Portugal	44.8	Singapore	20.4
Jamaica	26.2	Liberia	159.2	Sweden	8.5	Taiwan	19.1
Mexico	60.9	Morocco	149.0	Switzerland	12.8	Trinidad	26.2
Nicaragua	46.0	Burundi	150.0	Britain	17.5	Jordan	21.3
Panama	34.1	Central African Republic	190.0	Greece	27.8	South Korea	58.0
Peru	65.1	Chad	160.0	Spain	15.1	Papua	10.2
Uruguay	40.4	Dahomey	109.6	Yugoslavia	43.3	Philippines	67.9
Bolivia	60.4	Ethiopia	84.2			Syria	21.7
El Salvador	58.2	Guinea	216.0			Thailand	27.0
Honduras	39.3	Kenya	55.0			Turkey	153.0
Paraguay	38.6	Madagascar	102.0			Vietnam	100.0
Haiti	NA	Malawi	148.3			Afghanistan	400.0
		Mali	120.0			Bangladesh	124.3
		Mauritania	187.0			Burma	200.0
		Niger	200.0			Cambodia	100.0
		Rwanda	132.9			India	60.6
		Sierra Leone	170.0			Laos	NA
		Somalia	158.0			Nepal	NA
		Sudan	129.4			Pakistan	124.3
		Tanzania	162.5			Sri Lanka	45.1
		Togo	127.0			South Yemen	80.0
		Uganda	160.0			Yemen	50.0
		Upper Volta	180.0				
		Zaire	104.0				

spreads across the batches. In this section we discuss the construction of simple graphic displays for multiple unordered batches and describe procedures for finding useful spread-stabilizing transformations.

Parallel Graphic Displays

Parallel stem-and-leaf displays and parallel schematic plots permit ready contrasts and comparisons for studying location, extremes, and spreads in batches, and they facilitate the informal evaluation of distribution differences. Their construction is also quite easy. First consider a parallel stem-and-leaf display. This display consists of several sets of stems placed side by side with a common set of starting parts.[8] To assure that the display's starting parts are applicable for each batch, one determines the minimum and maximum and range of the entire data set. Previously described construction rules are then followed for each individual batch. Consider by way of example the infant mortality rates. Table 3 displays a set of multiple unordered batches where the nations have been placed into one of four geographical regions: the Americas, Africa, Europe, and Asia-Oceania. Note that their geographical classification is here considered to be purely qualitative.

We noted the range of the entire data set—from 9.6 for Sweden to 650.0 for Saudi Arabia—and chose 10^0 as a unit for the display. We then created one set of starting parts, from 9 to 20 with a HI stem, and formed four separate displays side by side, one for each batch, using only the one set of starting parts. Figure 8 is the completed parallel stem-and-leaf display for this data set. The display is quite revealing, highlighting the differences in spread and shape of the rates in the different batches. European nations are all very similar with quite low infant mortality rates. American nations are less similar as a batch than the European nations, with a spread twice as large and larger rates on average. But note the very large spreads of the African and Asian-Oceanian batches of nations.

[8]When only two batches are involved, the common starting parts may be placed between the batches with the leaves of one extending to the left and those of the other to the right of the starting parts.

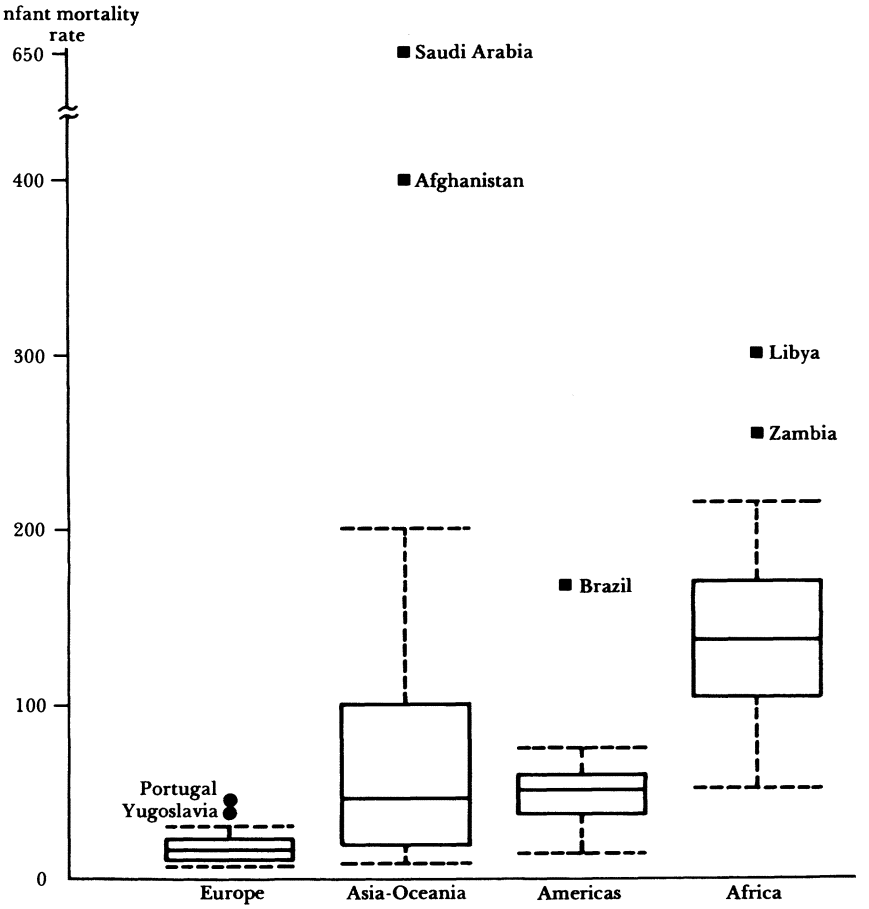
Figure 8. Parallel stem-and-leaf display of infant mortality rates for nations of the world classified by region.

	Americas	Africa	Europe	Asia-Oceania
0			9	
1	67		7302711275	616390
2	6		3057	8206117
3	498			2
4	860		43	5
5	1948	85		08
6	2050	3		70
7	889	16		
8		64		0
9				
10		924		00
11		4		
12		097		544
13		782		
14		98		
15		908		3
16		020		
17	0	0		
18		070		
19		0		
20		0		0
HI		300.0, 259.0, 216.0		650.0, 400.0

NOTE: Unit = 10°.

A parallel schematic plot can be drawn without difficulty when information contained in the parallel stem-and-leaf display is used. Simply find the order statistics comprising the letter-value display and the adjacent values for each of the batches. Using only one vertical scale for the entire set of batches, draw one schematic for each batch based on the values in each letter-value display. These are placed side by side (or one above the other with schematics lying horizontally). Figure 9 shows a parallel schematic plot for the infant mortality rates, one plot for each of the four batches of nations. Now we can see how easy it is to compare the batches by employing the graphic display. The outliers (Portugal and Yugoslavia; Afghanistan and Saudi Arabia; Brazil; and Libya and Zambia) are clearly differentiated from the remaining data values, and the differences in location and spread across batches can be seen. It is also apparent that there is a tendency for the variation in a batch to increase as the average value in-

Figure 9. Schematic plots of infant mortality rates for nations classified by region.



creases. The median infant mortality rate increases as we compare Europe to the Americas to Africa, but so does the *H* spread. This apparent functional relationship between medians and *H* spreads is not desirable, since it makes cross-batch comparison of location difficult. This situation can usually be corrected through the use of a suitable transformation of the data.

Transformations for Stabilization of Spread

We have seen that transformations are very helpful in the analysis of single batches since it is much easier to summarize

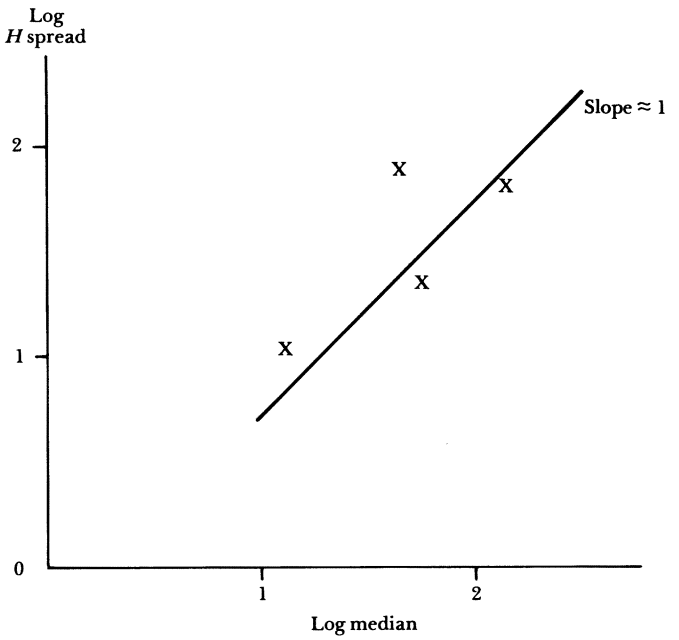
a batch that is symmetric and possibly well behaved. When studying multiple batches, one finds it easier to compare locations if spread is relatively constant across the batches.

The most important characteristic that the analyst examines when comparing batches is the location or average value of each batch. When spreads vary greatly across the batches, however, it may be difficult, and perhaps even impossible, to determine whether the apparent differences in average value are due to actual differences in location or to differences in spread. Consider again the schematic plots of infant mortality rates by nations classified by geographical region in Figure 9. There seem to be higher rates in the Americas than in Asia-Oceania, indicated by the difference between median values. But because the spreads differ greatly, this conclusion would be premature. Indeed, there are many nations in Asia-Oceania having rates that are higher than the maximum rate in the Americas. If only spreads are to be compared, in which case location is irrelevant, then the task is greatly simplified. The opposite statement is not true.

Mathematically, an analysis is difficult when the spread of a batch is functionally related to the magnitude of the median. In the infant mortality example, there is a tendency for spreads to increase as medians increase. We seek a transformation that will stabilize spread from batch to batch. Symmetry of spread within single batches is helpful for summarizing a batch, but equalization of spread across batches is essential for many comparisons. Lack of equality is regularly encountered in practice. In a data set of paired observations, values of one variable often increase in spread as the other variable increases. This heteroscedasticity is a plague of many empirical regression analyses. In the classroom, exposure to transformations to stabilize spread in a set of multiple unordered batches provides students with a simple initiation to a problem likely to arise in complex situations.

Tukey suggests remedying the problem of location-spread interaction by a power transformation of the data. But unlike the theory for variance-stabilizing transformations of Curtiss (1943) and Bartlett (1947), one need not make any distributional assumptions specifying the exact functional relationship between the median (M) and the H spread for the EDA procedure. Tukey's procedure is simply to plot log H spread against log M, one point for each batch. If the resulting scatterplot is roughly linear with

Figure 10. A (log H spread, log M) plot for transformation of infant mortality batches.



slope b, then $p = 1 - b$ is the appropriate power for a transformation of the data. Figure 10 is such a plot for the four batches of infant mortality rates. The slope of the line is nearly 1.0 so that $1 - 1.0 = 0.0$, or logarithms, is the "correct" transformation for these data.

So far, Tukey has not provided a theoretical justification for his recommendation of $p = 1 - b$ as the power for transformation. Our justification follows and is based on reasoning similar to that behind variance-stabilizing transformations.

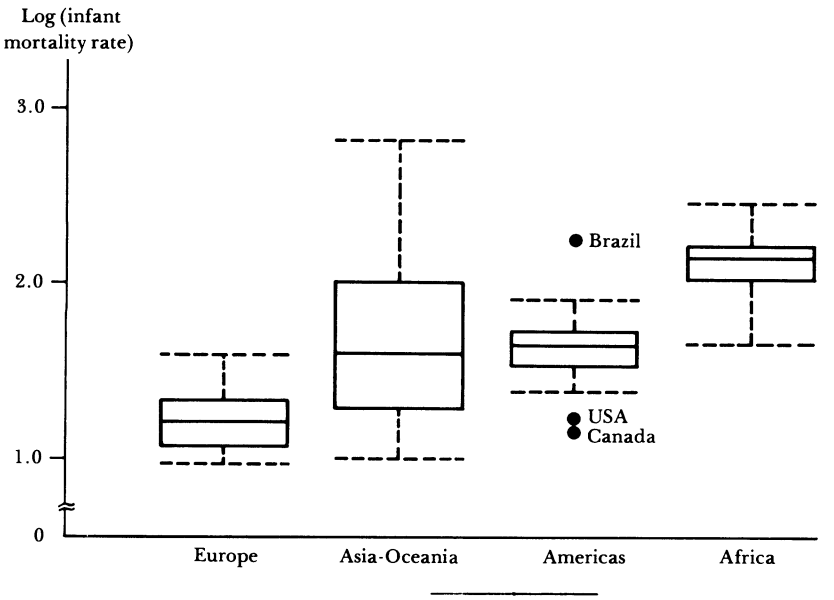
Assume that the spread of the batch (we shall use s as some measure of spread) is functionally related to the location of the batch, where $\tilde{y}$ is some measure of location; that is,

$$s_y^2 = f(\tilde{y}) \tag{16}$$

If $s_y = \sigma_y$, the standard deviation of the random variable Y, then

$$\begin{aligned} \sigma_g^2 &\approx (dg(\tilde{y})/d\tilde{y})^2 \sigma_y^2 \\ &\approx (dg(\tilde{y})/d\tilde{y})^2 f(\tilde{y}) \end{aligned} \tag{17}$$

Figure 11. Logarithms of infant mortality rates for nations of the world classified by region.



by a Taylor series expansion, where $g(Y)$ is some transformation of Y. We would like σ_g^2, the variance of the new random variable $g(Y)$, to be constant so that on the new g scale there will be no location-spread interaction. Thus we desire

$$\sigma_g^2 = C$$

where C is any constant. After some calculations, we find that

$$g(Y) = \int C/\sqrt{f(\tilde{y})}\, dy \tag{18}$$

is the "correct" variance-stabilizing transformation. If $f(\tilde{y}) = \tilde{y}$, for example, as with Poisson random variables, then $g(Y) = \sqrt{Y}$.

The EDA procedure for finding a transformation that stabilizes spread considers transformations of the form $g(Y) = Y^p$. Moreover, we assume that if $s_y = H$ spread, the approximate interquartile distance, then

$$\log s_y = a + b \log \tilde{y} \tag{19}$$

where $\tilde{y}$ = median. Thus, by applying Equation (17),

$$\begin{aligned} s_p^2 &\approx [(d/d\tilde{y})\, \tilde{y}^p]^2\, s_y^2 \\ &\approx [(d/d\tilde{y})\, \tilde{y}^p]^2\, e^{a+b(\log \tilde{y})} = C \end{aligned} \tag{20}$$

where C is the constant H spread for the transformed data Y^p.

Upon simplifying (20),

$$\begin{aligned} \tilde{y}^p &= C' \int e^{-b \log \tilde{y}}\, d\tilde{y} \\ &= C' \int \tilde{y}^{-b}\, d\tilde{y} = C'' \tilde{y}^{(1-b)} \end{aligned} \tag{21}$$

where C' and C'' are constants. Hence $\tilde{y}^p = C''\tilde{y}^{(1-b)}$, and therefore $p = 1 - b$ is the power transformation for the data that stabilizes spread, assuming (19) to be true.

A logarithmic transformation for the infant mortality multiple batches is very effective. Figure 11 shows the schematic plots for the transformed data. Comparisons are now much easier to make, since the spreads of all batches except the Asia-Oceania batch are nearly equal. There appears to be a difference in average infant mortality rate across batches.

TOOLS FOR ORDERED MULTIPLE BATCHES AND (X, Y) DATA

In the preceding section we discussed exploratory data analysis procedures for a set of unordered multiple batches—unordered in a quantitative sense. But now suppose we have a collection of batches that *can* be ordered along some natural scale. The batches in this ordered collection are related to each other *quantitatively*. Each batch can be assigned a value of some ordering variable, and comparisons of batches along this new dimension are possible. Examples include crime rates for cities classified by population (population is the ordering dimension and, presumably, cities would fall into one of several population classes); net interest costs for bond sales for public schools, each bond issue classified by its Moody rating (an assumedly simple unit-interval scale from Aa, very little risk, to Ba, slight risk, to C, substantial risk); and infant mortality rates for nations classified into one of four per capita income groups (highest, higher, middle, lower per capita income, each bounded by specified incomes). The analysis of such data sets focuses on the relation between the batches and the ordering dimension.

The experienced analyst may note that a collection of ordered multiple batches is actually a set of paired observations: Each data value in batch i, Y_{ij}, is assigned a value X_i from the

ordering variable, resulting in a set $\{(X_i, Y_{ij}),\ i = 1, 2, \ldots, I;\ j = 1, 2, \ldots, n_i\}$ of paired observations where I = number of batches and n_i = number of observations in batch i. Hence a collection of ordered multiple batches may be analyzed as a set of (X, Y) data by fitting lines to summarize the relationship between X and Y and then examining residuals. But by doing so one ignores substantial information, primarily on the relationships between the batches. The emphasis in an analysis of an (X, Y) data set is on the summarization of the data by a linear equation, and not on how the data associated with one or more specific X values vary. The EDA procedures discussed in this section are sufficiently simple and resistant that we often prefer to explore a data set of paired observations by reconstructing them as multiple batches. We discuss this and other procedures for analyzing an (X, Y) data set, such as fitting resistant lines, in this section.

Conditional Typical Values for Summarization

Suppose we have a collection of several batches that can be ordered along some numerical variable or scale. We could begin our exploratory analysis of this data set as if the collection were unordered by constructing parallel stem-and-leaf displays and parallel schematic plots for the batches, as discussed in the preceding section. These displays give us a good feel for the variability of the data. When constructing the parallel schematic plots, however, the X axis now becomes important. This horizontal axis corresponds to the ordering variable of the data set. The position of the schematic plots along the X axis corresponds to the value of the ordering variable assigned to each batch.

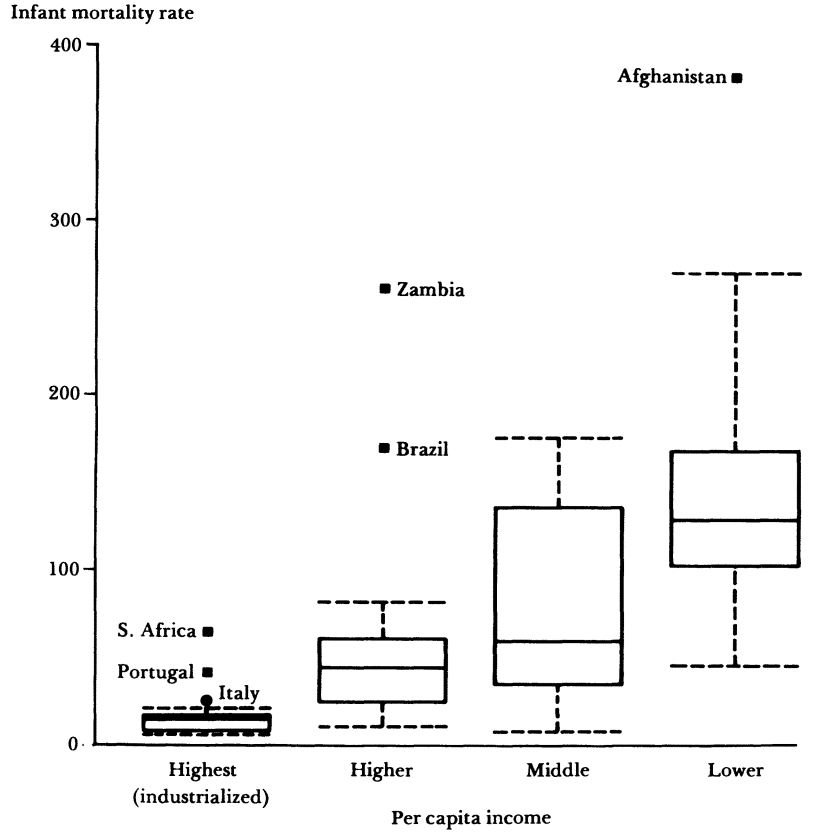
An example will serve to illustrate these points. Consider again the infant mortality rates for nations of the world, but now classify the nations according to their per capita income. As the data were originally organized (in *The New York Times*) there were five levels of per capita income: highest (for industrialized nations), high, middle, lowest, and petroleum exporters. We have purposely left out the nine nations that *The New York Times* categorized as petroleum exporters, since per capita income is not a good measure of the wealth of these countries and consequently their position as a group along the ordering dimension is not informative. The data are shown in Table 4.

TABLE 4

Infant Mortality Rates for Nations of the World Classified by Per Capita Income

Highest (Industrialized)		Higher		Middle		Lowest	
Australia	16.7	Argentina	59.6	Bolivia	60.4	Afghanistan	400.0
Austria	23.7	Brazil	170.0	Cameroon	137.0	Bangladesh	124.3
Belgium	17.0	Chile	78.0	Congo	180.0	Burma	200.0
Canada	16.8	Colombia	62.8	Egypt	114.0	Burundi	150.0
Denmark	13.5	Costa Rica	54.4	El Salvador	58.2	Cambodia	100.0
Finland	10.1	Dominican Republic	48.8	Ghana	63.7	Central African Republic	190.0
France	12.9	Greece	27.8	Honduras	39.3	Chad	160.0
West Germany	20.4	Guatemala	79.1	Ivory Coast	138.0	Dahomey	109.6
Ireland	17.8	Israel	22.1	Jordan	21.3	Ethiopia	84.2
Italy	25.7	Jamaica	26.2	South Korea	58.0	Guinea	216.0
Japan	11.7	Lebanon	13.6	Liberia	159.2	Haiti	NA
Netherlands	11.6	Malaysia	32.0	Morocco	149.0	India	60.6
New Zealand	16.2	Mexico	60.9	Papua	10.2	Kenya	55.0
Norway	11.3	Nicaragua	46.0	Paraguay	38.6	Laos	NA
Portugal	44.8	Panama	34.1	Philippines	67.9	Madagascar	102.0
South Africa	71.5	Peru	65.1	Syria	21.7	Malawi	148.3
Sweden	9.6	Singapore	20.4	Thailand	27.0	Mali	120.0
Switzerland	12.8	Spain	15.1	Turkey	153.0	Mauritania	187.0
Britain	17.5	Taiwan	19.1	Vietnam	100.0	Nepal	NA
United States	17.6	Trinidad	26.2			Niger	200.0
		Tunisia	76.3			Pakistan	124.3
		Uruguay	40.4			Rwanda	132.9
		Yugoslavia	43.3			Sierra Leone	170.0
		Zambia	259.0			Somalia	158.0
						Sri Lanka	45.1
						Sudan	129.4
						Tanzania	162.5
						Togo	127.0
						Uganda	160.0
						Upper Volta	180.0
						South Yemen	80.0
						Yemen	50.0
						Zaire	104.0

Figure 12. Schematic plots of infant mortality rates for nations classified by income.



The ordering variable is of course per capita income, and for illustrative purposes we space the four schematic plots at equal intervals along the X axis of the display in reverse order, highest first and lowest last. As can be seen in Figure 12, this reverse ordering highlights the lack of constant spread in the data set: As per capita income decreases, the infant mortality rate increases, but so does the variation in the infant mortality rates. The data show an obvious location-by-spread interaction. Transformations to alleviate this situation can be more difficult to find, since we also must consider the effect of a transformation to stabilize spread on the relationship between the location of the batches and the X ordering variable. We shall come back to this problem later.

There are two important issues in the analysis of a collection of ordered batches: (1) How can we effectively summarize the

information within a batch? (2) What is the relationship between this summary information and the X ordering variable? We find it useful (Leinhardt and Wasserman, 1978; in press) to work with a concept we call the conditional typical value and use it to summarize the information in a collection. Here the conditional typical value of a batch is merely the median value of the batch. The second issue is concerned with the relationship between the conditional typical values and the X variable of the data set: Is the relationship linear? And if not, can it be made linear by a transformation? We shall discuss this second issue in the next section.

A conditional typical value is a summary statistic, representative of a specific batch. If y is any number within a single batch of numbers, for example, then

$$C(y \mid y \in \text{batch of numbers}) = \text{median of batch} \qquad (22)$$

is the conditional typical value of the batch. As noted in an earlier section, $C(y \mid y \in \text{batch})$ is merely the median of the batch. Suppose we have a set of batches, either ordered or not: $\{\text{batch}_1, \text{batch}_2, \ldots, \text{batch}_I\}$. Then, if y is any number,

$$C(y \mid y \in \text{batch}_i) = \text{median of batch}_i \qquad (23)$$

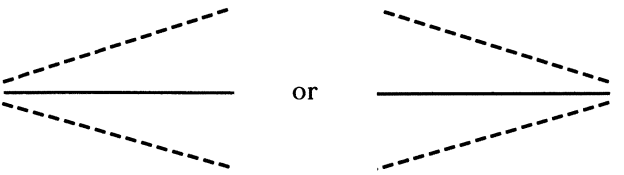
for $i = 1, 2, \ldots, I$. Hence we can reduce the information in a collection of I batches to I representative conditional typical values.

Searching for Linearity

Conditional typical values, defined in Equation (23), are good estimators of the values within the batches. We can study how these summary statistics change as we move from batch to batch. Do they increase or decrease as the X ordering variable increases? And if so, is the increase or decrease roughly linear, or is the functional relationship of some higher degree? These are important questions, primarily because the relationship between the data—as reflected by the summary conditional typical values and the X variable—is our second fundamental issue as mentioned in the previous section. If the relationship is linear, or can be made so by a suitable algebraic transformation on the data, then our summarization task is simplified because such data can be readily summarized by a linear equation. In this section we discuss an EDA diagnostic display that is useful in searching for linearity.

A scatterplot of the conditional typical values versus the corresponding values of the X variable is quite informative. If we connect these batch medians by line segments, we can study how linear this (X, Y) relationship actually is. This jagged line is called a *median trace*. Tukey also suggests drawing *hinge* traces, one for the set of upper hinges and another for the lower hinges, where hinges for batches are connected by line segments. The hinge traces not only shed light on the linearity of the relationship but they also allow one to study the change in spread of the data as X increases. (Other traces, slicing the data at different locations, are also possible and sometimes useful.)

If the relationship were truly linear, then the hinge traces would have the same slope as the median trace, forming three parallel lines. Unequal slopes usually imply a lack of linearity unless the traces have the characteristic megaphone or wedge shapes:



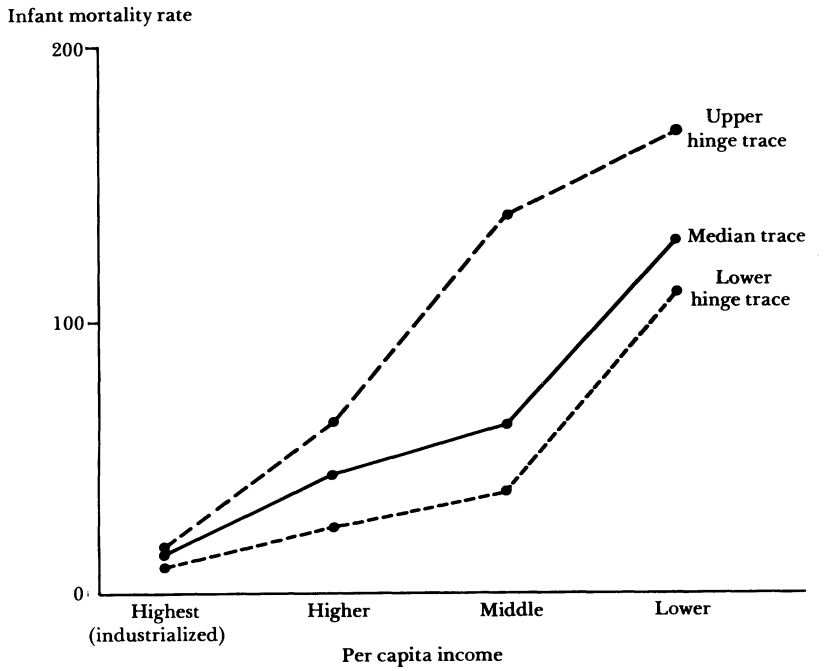
These shapes indicate that the spread of the batch values increases or decreases as we move from batch to batch. In such situations the batch medians do not summarize the data well. Figure 13 shows the median and hinge traces for the infant mortality data. These data exhibit both a slight nonlinearity and a definite increase in spread for decreasing per capita income.

As can be seen, the median and hinge traces are useful diagnostic tools. The search for an appropriate transformation must not only stabilize spread but also promote linearity between the data and the X variable. We discuss procedures for finding such transformations in the next sections.

Transformations to Stabilize Spread

We have seen by examination of the plot of hinge and median traces that there may be two different reasons for transforming a set of ordered multiple batches: (1) to stabilize the

Figure 13. Median and hinge traces for infant mortality rates.



spread of the data values and (2) to achieve a linear relationship between the data and the X variable. Empirically, these two goals are seldom disparate, as we shall attempt to show.

If the nonlinearity is monotone, linearity is usually increased by transforming the X variable. Transforming X to higher powers has the effect of expanding the X axis and promotes linearity in plots that resemble e^x or $-e^x$ exponential functions. Transforming X to smaller powers has the effect of shrinking the X axis and promotes linearity in plots that resemble the negative exponential functions e^{-x} or $-e^{-x}$. For the infant mortality data, as per capita income decreases, infant mortality rate appears to grow exponentially, so that the traces are roughly e^{-x} functions. A transformation of X down the ladder of powers seems necessary.

Spread is often stabilized by transforming the data values themselves, as with multiple unordered batches. A log median versus log hinge spread plot is occasionally helpful. We should note that a transformation on X to promote additivity will usually

decrease differences in spread, and a transformation on Y to stabilize spread may make the traces more linear.

As an aside: A set of multiple batches, either unordered or ordered, is usually analyzed by a one-way analysis of variance. One computes an effect for each group or batch, determines whether the effects differ in any way by an F test, and if so, tries to find out how they differ. An exploratory analysis of the same data set does not involve any hypothesis testing, either because the analyst may not believe the ANOVA assumptions are justified, or merely because he or she simply does not want to make any confirmatory tests during the early stages of the analysis. However, the ANOVA notion of an "effect" has an EDA counterpart.

In a one-way ANOVA, the effect α_i of group or treatment i is estimated by

$$\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}, \quad i = 1, 2, \ldots, I; \ \Sigma\hat{\alpha}_i = 0 \tag{24}$$

where $\bar{y}_{i\cdot}$ is the mean of the ith group and $\bar{y}_{\cdot\cdot}$ is the grand (overall) mean of the data. If we compute y_M, the grand median of the data, then

$$\hat{\gamma}_i = C(y \mid y \in \text{batch}_i) - y_M, \quad i = 1, 2, \ldots, I \tag{25}$$

also estimates α_i, computed with conditional typical values, such that median $(\hat{\gamma}_i)$ is roughly zero. Thus we can define a set of treatment effects, as in ANOVA, that are centered at zero but have a certain amount of resistance to stray or outlying values. Moreover,

$$\hat{\sigma} = \tfrac{3}{4} \operatorname*{median}_i \ (H\,\text{spread}_i) \tag{26}$$

where $H\,\text{spread}_i$, the H spread or interquartile distance of batch$_i$, is a resistant estimate of the within-group or within-batch standard error and provides a standard for comparing batches.

Linear Fits and Linearizing Transformations for (X, Y) Data

The first part of this section has been concerned solely with data sets of ordered multiple batches. We now extend the EDA procedures described for ordered multiple batches to sets of paired (X, Y) observations.

We have shown in the introduction to this section how a set of multiple ordered batches can be viewed as an (X, Y) data set. Moving in the reverse direction, constructing a set of ordered multiple batches from an (X, Y) data set, is straightforward. We examine the X variable and use it as the X-ordering dimension for a set of multiple ordered batches. Thus we break X up into m intervals $(X_{(0)}, X_{(1)})$, $(X_{(1)}, X_{(2)}), \ldots, (X_{(m-1)}, X_{(m)})$, where $X_{(0)} = \min(X_i)$ and $X_{(m)} = \max(X_i)$, such that:

1. The widths of the intervals $X_{(i+1)} - X_{(i)}$ are roughly equal.
2. The number of Y_i's falling within the intervals are roughly equal.

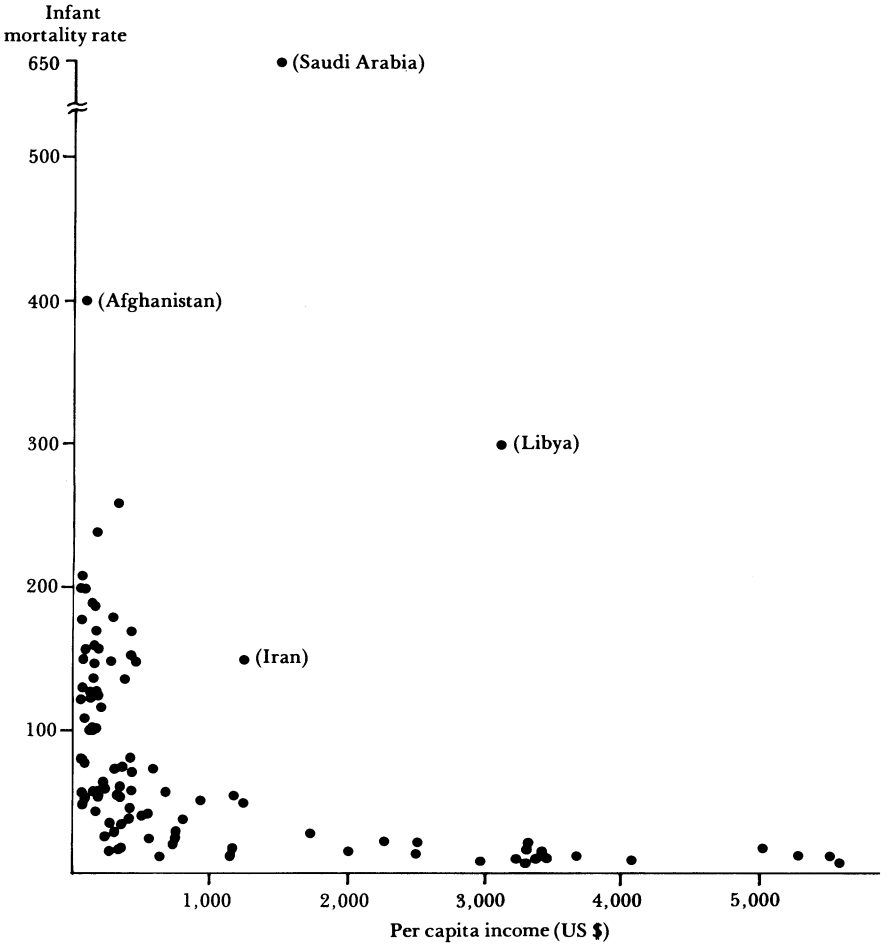
We could also use the letter values of X for breaking X up into intervals, as suggested by Tukey (1977). We may not be able to achieve all three goals simultaneously, but we strive for an optimal solution. We group together all the Y's with paired X's between $X_{(i)}$ and $X_{(i+1)}$ as the ith batch and consequently obtain m ordered batches.

As an example, consider again the infant mortality rates for nations. We now examine the per capita incomes for each of the countries in 1970 in US dollars, so that we have a set of 101 paired observations (including the petroleum exporters). The scatterplot of these data in Figure 14 shows the same nonlinearity and heteroscedasticity as the hinge and median traces of the batches shown in Figure 13. The set of four multiple batches of infant mortality rates analyzed earlier was constructed from the paired data, defining the X intervals as (0, 125), (125, 400), (400, 2,500), (2,500, 6,000).

The primary reason for making an (X, Y) data set into a set of m ordered multiple batches is to obtain a situation that facilitates insight into the nature of the data. Heteroscedasticity easily becomes apparent if the batches exhibit unequal spreads, and nonlinearity is quickly revealed from the hinge and median traces. In short, the analysis becomes more exploratory. Consequently, an EDA procedure for fitting lines to (X, Y) data uses the data set in the multiple batch form, as we shall see shortly.

Linear fits, with or without a transformation of the data, are very useful and interpretable. A good-fitting line with small

Figure 14. Scatterplot of infant mortality rate versus per capita income for nations.



residuals is an excellent summarization device—we merely need to know the slope and intercept of the line[9] and the linearizing transformations[10] for both X and Y: four pieces of information. This is in contrast to m pairs of conditional typical values and corresponding values of the X-ordering variable that are needed just to

[9]Which Tukey calls "flattening" the data—subtracting out the line to leave residuals.

[10]Which Tukey calls "straightening" the data—making the relationship between X and Y linear.

begin to summarize the information contained in a set of ordered multiple batches—or perhaps even the entire (X, Y) data set if no linear relationship can be found.

A basic EDA line-fitting procedure utilizes the (median X, median Y) value from each of the multiple batches constructed from the paired data. Note for the ith batch that (median X, median Y) $\approx$ (midpoint of ith X interval, conditional typical value of batch$_i$). Hence conditional typical values play an important role in summarizing (X, Y) data also.

We choose three (median X, median Y) points, one from the batch associated with the smallest or first third of the values of X, one from a "middle" batch or second third, and one from the batch associated with the largest or third third of the values of X. These three summary points, $(X_{(1)}, Y_{(1)})$, $(X_{(2)}, Y_{(2)})$, $(X_{(3)}, Y_{(3)})$, in order of increasing X, are used first to determine how linear the data are and second to compute a fitted line for the data. Tukey calls the equation produced by this algorithm a *resistant line*.

We compute

$$\begin{aligned} m_1 &= (Y_{(2)} - Y_{(1)})/(X_{(2)} - X_{(1)}) \\ m_2 &= (Y_{(3)} - Y_{(2)})/(X_{(3)} - X_{(2)}) \end{aligned} \tag{27}$$

which are the slopes of the lines connecting $(X_{(1)}, Y_{(1)})$ and $(X_{(2)}, Y_{(2)})$ and $(X_{(2)}, Y_{(2)})$ and $(X_{(3)}, Y_{(3)})$. If the data are nonlinear, $m_1 \neq m_2$ and we begin an exploratory search for power transformations of X and Y such that

$$\begin{aligned} m_1(p, q) &= (Y^q_{(2)} - Y^q_{(1)})/(X^p_{(2)} - X^p_{(1)}) \\ &\approx m_2(p, q) = (Y^q_{(3)} - Y^q_{(2)})/(X^p_{(3)} - X^p_{(2)}) \end{aligned} \tag{28}$$

That is, the slopes of the connecting line segments become roughly equal.

Tukey's procedure for finding p and q is purely exploratory. Using it involves working with the three summary points until $m_1(p, q)$ approximately equals $m_2(p, q)$. We have developed the following procedure for finding a power p for X, while holding q fixed at unity, that equates $m_1(p, 1)$ and $m_2(p, 1)$. The computation involves a Taylor series expansion of X^p, as did the derivation of the procedure for finding a symmetrizing transformation for single batches.

We seek to transform X such that

$$\begin{aligned} m_1(p, 1) &= (Y_{(2)} - Y_{(1)})/(X_{(2)}^p - X_{(1)}^p) \\ &= m_2(p, 1) = (Y_{(3)} - Y_{(2)})/(X_{(3)}^p - X_{(2)}^p) = b \end{aligned} \tag{29}$$

where $Y = a + bX^p$ is the fitted resistant line. Note that this criterion (29) is equivalent to either forcing the angle Θ between the line segment connecting $(X_{(1)}, Y_{(1)})$ and $(X_{(2)}, Y_{(2)})$ and the line segment connecting $(X_{(2)}, Y_{(2)})$ and $(X_{(3)}, Y_{(3)})$ to be 180° or the tangent of this angle Θ to be tan 180° = 0:

$$\begin{aligned} &\tan 180° \\ &\quad = [m_2(p, 1) - m_1(p, 1)]/[1 + m_1(p, 1)m_2(p, 1)] = 0 \end{aligned} \tag{30}$$

We use a Taylor series expansion for X^p, so that the criterion (29) becomes: Find p such that

$$\begin{aligned} &(X_L/X_H)^{(p-1)} \\ &\quad = (Y_{(2)} - Y_{(1)})/(X_{(2)} - X_{(1)}) \cdot (X_{(3)} - X_{(2)})/(Y_{(3)} - Y_{(2)}) \end{aligned} \tag{31}$$

where X_L lies between $X_{(1)}$ and $X_{(2)}$ and X_H between $X_{(2)}$ and $X_{(3)}$. We find that

$$p_0 = 1 + [\log m_1(1, 1) - \log m_2(1, 1)]/[(\log X_L - \log X_H)] \tag{32}$$

is the appropriate power for transformation, where X_L and X_H are positive. Note that this procedure is only appropriate when $m_1(1, 1)$ and $m_2(1, 1)$ have the same sign.

For the infant mortality rates-per capita income data set, the three summary points are

$$\begin{aligned} (X_1, Y_1) &= (100, 125) \\ (X_2, Y_2) &= (334, 58) \\ (X_3, Y_3) &= (2,526, 20.4) \end{aligned}$$

and

$$m_1(1, 1) = -0.286 \qquad m_2(1, 1) = -0.0172$$

and

$$X_L = 217 \qquad X_H = 1{,}430$$

that is, $(X_{(1)} + X_{(2)})/2$ and $(X_{(2)} + X_{(3)})/2$, respectively. Hence

$$\begin{aligned} p_0 &= 1 + [\log (0.29) - \log (0.02)]/[\log (217) - \log (1{,}430)] \\ &= -0.42 \end{aligned}$$

If we take negative reciprocal roots of X—that is, choose $p_0 = -0.5$—we have

$$m_1(-0.5, 1) = -1{,}488.9 \qquad m_2(-0.5, 1) = -1{,}074.3$$

which is an improvement—with a ratio $m_2/m_1 = 0.72$ contrasted with an original ratio of 0.06.

Because the slopes are still not quite equal, transformations on Y *and* X may also be sought. After some exploration, we find that transforming to log X and log Y also straightens the data and has a more intuitively appealing meaning. With $p = 0$ *and* $q = 0$,

$$m_1(0, 0) = -0.64 \qquad m_2(0, 0) = -0.51$$

which has a ratio $m_2/m_1 = 0.80$. The scatterplot of log (infant mortality rate) versus log (per capita income) in Figure 15 shows a fair linear trend.[11]

Once we have found a pair (p_0, q_0) of linearizing powers for transformation, we let

$$b = (Y_{(3)}^{q_0} - Y_{(1)}^{q_0})/(X_{(3)}^{p_0} - X_{(1)}^{p_0}) \tag{33}$$

and

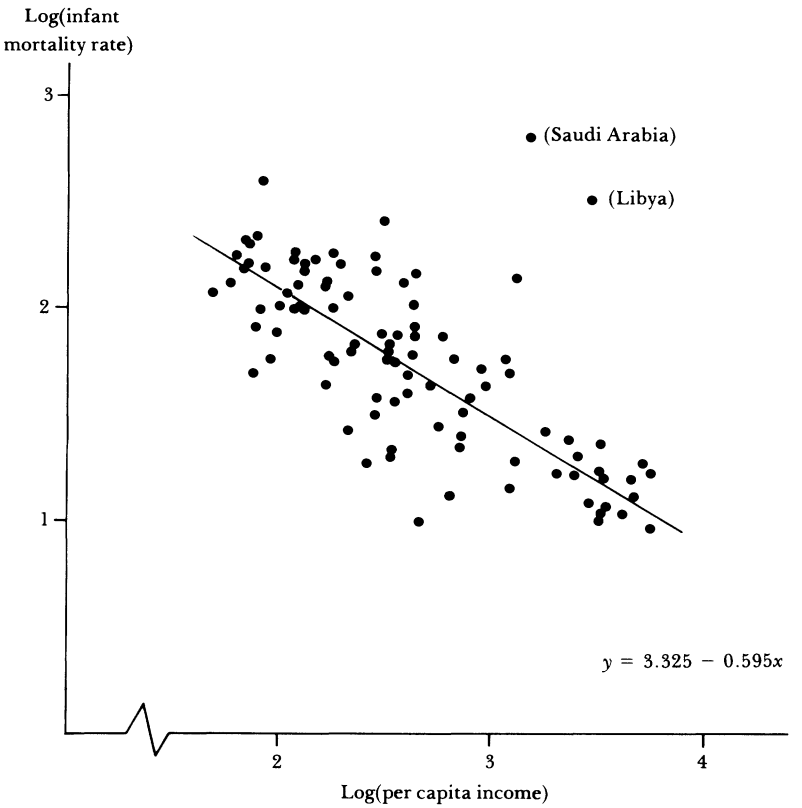
$$a = \tfrac{1}{3}[(Y_{(1)}^{q_0} - bX_{(1)}^{p_0}) + (Y_{(2)}^{q_0} - bX_{(2)}^{p_0}) + (Y_{(3)}^{q_0} - bX_{(3)}^{p_0})] \tag{34}$$

be the slope and intercept, respectively, of the resistant line relating Y to X. For the infant mortality data, the equation is log $(Y) = 3.205 - 0.561$ log (X)—that is, a 1 percent increase in per capita income results in a 0.561 percent decrease in the infant mortality rate since 0.561 is the computed "elasticity" of infant mortality with respect to income.

After computing this initial fit, we find the batch of residuals $\{Y_i - a - bX_i\}$ and plot them against X. Often some trend is still apparent in this plot, and thus we must adjust a and b slightly. This routine, which Tukey calls *polishing,* is an iterative process aimed at producing a set of residuals that shows no relation to X. The process is straightforward. After computing $r_i^{(1)} = Y_i^{q_0} - a - bX_i^{p_0}$, we fit a resistant line for $r_i^{(1)}$ on $X_i^{p_0}$, $r^{(1)} = a^{(1)} +$

[11] The transformation on Y may be interpreted as spread-stabilizing, as before, and that on X as linearizing.

Figure 15. Scatterplot of log (infant mortality rate) versus log (per capita income) for nations, with resistant line.



$b^{(1)}X^{p0}$, and add this fit to the initial fit of Y^{q0} on X^{p0}. This new resistant line is $Y^{q0} = (a + a^{(1)}) + (b + b^{(1)})X^{p0}$. We keep polishing until $a^{(i)}$ and $b^{(i)}$ computed on the ith iteration are essentially zero.[12] For the infant mortality data, the polished line is $\log(Y) = 3.325 - 0.595 \log(X)$ as plotted on Figure 15. Thus fitting resistant lines is an iterative process in which we continue pulling out parts of a fit from the residuals of the previous iteration, adding this equation to the initial equation, until no fit is left in the resid-

[12]When using a computer, it is advantageous to limit arbitrarily the number of iterations to avoid trivial changes that result from machine precision.

uals. One can think of this as a process in which, in the equation data = fit + residual, part of the fit is transferred at each iteration from residual to fit.

After finding the polished fit, $Y^{q0} = a^* + b^* X^{p0}$, residuals should be computed and analyzed as a single batch of data. All too often the fit component of the equation data = fit + residual is stressed and the residual component ignored. Residuals as a batch should appear well behaved, should be centered at zero, and should be a random swarm when plotted against X^{p0}. Once the major linear (after straightening through some transformation) trend has been removed and the data flattened to zero, analysis of the residuals may reveal other features of the data such as periodicities.

We should comment on the relation of a resistant line to a least-squares line. A least-squares fit is very sensitive to outlying points. This is not true of a resistant line. If the data are relatively linear, the spread about the line is constant, and there are no outliers, then the least-squares line will coincide with the resistant line. If these conditions are not met, the two lines will differ, with the resistant line usually being a better summary of the data than the least-squares line.

TOOLS FOR TWO-WAY TABLES

We now consider the analysis of a two-way table,[13] a two-dimensional array of *responses* to the combinations of two factors, one factor associated with the rows of the array and the other with the columns. Such a table is $R \times C$ in dimension, R levels (or versions) of factor 1 and C levels (or versions) of factor 2. Examples include a table of prices of R foodstuffs in C different regions of the country or average national infant mortality rates categorized by region of the world and income level.

Classically a two-way table is decomposed into a grand mean, row effects, column effects, possibly interactions, and, of course, residuals. It is analyzed by examining ratios of sums of

[13] Tukey (1977) presents an extension to three-way tables that we do not discuss here.

squares in a two-way analysis of variance. The effects are all computed via row, column, and grand means. Assuming only one observation per cell, the model is

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad i = 1, 2, \ldots, R;\ j = 1, 2, \ldots, C \quad (35)$$

where $\hat{\mu}$ = grand mean = $\bar{y}_{\cdot\cdot}$

$\hat{\alpha}_i = i$th row effect = $\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$ $\quad i = 1, 2, \ldots, R$

$\hat{\beta}_j = j$th column effect = $\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}$ $\quad j = 1, 2, \ldots, C$

$e_{ij} = (i, j)$th residual = $y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot}$

are estimates of the parameters. The dot subscript indicates that the data have been averaged over that dimension. The errors ϵ_{ij} are assumed to be Gaussian random variables, with zero mean and variance σ^2, so that the sums of squares for the row and column effects, as well as the sum of squares for error, are multiples of χ^2 random variables. Consequently F ratios can be formed to test whether the levels of either factor are different from zero. More than one observation per cell permits estimating and testing interactions.

The EDA procedure is quite similar except that no distributional assumption is made for the errors and consequently no hypothesis tests are performed. Nonetheless, the resistant EDA measures give valuable insight into the presence or absence of row and column effects, help to determine whether interactions are present, and, if so, often suggest a power transformation that may eliminate them. Tukey's procedure, *median polish,* is a member of a class of techniques called *alpha polish.* Alpha polish is a minimization of the L_p-th norm of the residuals and, in general, requires a nonlinear optimization algorithm that cannot conveniently be solved by hand. However, there are two special alpha polishes—median polish and mean polish—in which solutions are readily obtained. We shall discuss the former and give examples.[14] Lastly, we present the theory behind *diagnostic plots,* an exploratory tool for determining a good transformation for a table to achieve additivity of the factor effects.[15]

[14]It should be noted that median polish is generally not appropriate when the data are counts in the form of a contingency table. However, transforming such data by taking logarithms of the cell counts and then using median polish yields an EDA analog to log linear models for discrete multivariate data.

[15]Tukey argues that additivity of effects, if an adequate summary of the data, is often a preferable representation than a more complex but possibly equally effective model involving multiplicative effects or interaction terms.

Calculating Fits with Alpha Polish

Consider the data given in Table 5. The array presents median infant mortality rates for the nations classified by region and per capita income. This is an example of a 4 × 4 two-way table of responses. For illustrative purposes we assume that there is only one response per cell. The general additive model for this table is

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \qquad i = 1,2,3,4;\ j = 1,2,3,4 \quad (36)$$

where the α's are the income level effects and the β's are regional effects.

The parameter estimates produced by two-way ANOVA are given after Equation (35). These estimates are the set of effects that minimize the sum

$$\sum_i \sum_j e_{ij}^2 = \sum_i \sum_j (y_{ij} - \mu - \alpha_i - \beta_j)^2 \quad (37)$$

That is, they are the least-squares estimates. However, the sum of squares (37) is unduly influenced by large residuals, as is true in regression. Hence we consider minimizing alternative quantities.

A procedure for finding reliable effect estimates is alpha polish. We assume Model (36) and find μ, $\{\alpha_i\}$, and $\{\beta_j\}$ by

$$\min_{\substack{\mu \\ \{\alpha_i\} \\ \{\beta_j\}}} (\Sigma\Sigma \mid e_{ij} \mid^p)^{1/p} = \min_{\substack{\mu \\ \{\alpha_i\} \\ \{\beta_j\}}} (\Sigma\Sigma \mid y_{ij} - \mu - \alpha_i - \beta_j \mid^p)^{1/p} \quad (38)$$

where $p = 2(1 - \alpha)$ for a fixed α, $0 \leq \alpha \leq 0.5$. Equation (38) is the pth norm of the residuals. When $\alpha = 0$, then $p = 2$ and we

TABLE 5
Median Infant Mortality Rates for Nations Categorized by Region and Per Capita Income

Per Capita Income Level	Americas	Africa	Europe	Asia-Oceania
Highest	17.2	71.5	15.3	16.2
High	15.7	167.7	27.8	20.4
Middle	48.8	137.0	NA	42.5
Lowest	NA	148.3	NA	100.0

have least squares and (38) is equivalent to (37). When $\alpha = 0.5$, then $p = 1$ and (38) gives the least-absolute-residuals solution.

Minimizing the quantity in Equation (38) produces parameter estimates that are resistant to large residuals, particularly when α is near 0.5. If $0 < \alpha < 0.5$, we cannot minimize (38) without the aid of a computer despite the intuitive appeal of intermediate α values. When $\alpha = 0.0$ or 0.5, however, parameter estimates are readily computed.

When $\alpha = 0.5$, Equation (38) becomes

$$\min \sum_i \sum_j | y_{ij} - \mu - \alpha_i - \beta_j | \tag{39}$$

a LAR (least absolute residuals) problem. Such minimization problems are called L_1 (first-norm) problems and are efficiently solved by linear programming (Barrodale and Roberts, 1974). Although a solution in this case usually requires a computer, median polish provides a useful and easily obtained approximation. It consists of iteratively subtracting the medians of the rows and columns of the table. The procedure involves first finding the row medians of the table and then subtracting from y_{ij} the median of the ith row.[16] Call these new values $\hat{y}_{ij}$. Next find the column medians of the table of $\hat{y}_{ij}$'s and subtract from $\hat{y}_{ij}$ the median of the jth column of the $\hat{y}_{ij}$'s. The algorithm continues, alternating the rows and columns, subtracting out medians each time. Eventually one obtains a table with zero row and column medians (assuming no missing values) and, by accumulating the extracted medians at each step, parameter or effect estimates. The values remaining in the table are residuals. Further details on the algorithm can be found in Tukey (1977). In this exploratory manner Singer (1976) analyzes a two-way table of the number of children born in six 10-year periods in the 1880s for 17 Norwegian counties.

For the nations data, the table of residuals with estimated effects along the border is given in Table 6. There are substantial differences among the effects: Africa has much higher average

[16] The medians extracted in the first iteration (or first and second) are used to obtain the grand median, "common," or "all" term. One can either start with the rows or the columns. The effects will not be unique, but they will not differ by very much.

TABLE 6
Median Polish (Effects and Residuals) of Median Infant Mortality Rates for Nations

Per Capita Income Level	Americas	Africa	Europe	Asia-Oceania	Income Effects
Highest	−1.77	−20.85	8.74	1.77	−29.91
High	8.74	45.36	−8.74	−24.02	0.08
Middle	0.00	14.81	NA	−1.77	−0.08
Lowest	NA	−14.81	NA	14.81	40.84
Regional effects	2.27	75.66	−10.14	−2.27	46.61 (grand median)

rates; Europe, much lower average rates. Low per capita income has a large effect, as does very high per capita income; but the cut we have made to define the middle and high income categories has not produced effects that are much different from one another within these categories. Note that estimates can now be provided for missing observations. For example, the estimated rate for low-income American nations is 89.72, row effect plus column effect plus grand median.

To evaluate how well a simple additive model fits the data, we should treat the residuals as a single batch and examine them in detail. Outliers should be noted and the data perused to discover the cause of large absolute residuals. Moreover, one should rearrange the "bordered table" so that the row effects increase from top to bottom and column effects increase from left to right.

Figure 16. Residuals from median polish fit to median infant mortality rates for nations.

```
−4 |
−3 |
−2 | 4 0
−1 | 1
−0 | 8 8 1 1
 0 | 0 8 1
 1 | 4 4
 2 |
 3 |
 4 | 5

 3 | NA
```

NOTE: Unit = 10^1.

Any evidence of an opposite-corner sign pattern in the residuals—positive residuals in the northwest and southeast corners, negative residuals in the northeast and southwest corners or vice versa—indicates a systematic lack of fit and the presence of an interaction between the factors. In the next section we discuss how to deal with such an interaction.

Figure 16 is a stem-and-leaf display of the residuals from a simple additive fit to the infant mortality data. We see one large positive outlier: 45.36, the (2, 2) residual. African nations in the high-income category have higher average mortality rates than the model predicts. The other residuals, although some are large, cluster reasonably well around zero.

Transformations to Obtain Additive Fits in Two-Way Tables

Sometimes a simple additive fit does not provide an adequate summary of data in a two-way table. Often this inadequacy results from the presence of an interaction between the two factors. When there are $n > 1$ observations per cell and interaction, the model

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \qquad k = 1, 2, \ldots, n \qquad (40)$$

provides a better summarization of the data. If we do have multiple observations per cell, we can test for the presence of the interaction terms $(\alpha\beta)_{ij}$ in Model (40) by standard ANOVA theory.

Frequently only one observation per cell is recorded, however; in fact, EDA procedures for two-way tables were devised for this type of data structure. In this case, there are no degrees of freedom for interaction in the ANOVA table. Tukey (1949) suggests an approximate one degree of freedom test for nonadditivity in a two-way ANOVA; a small p value for the F statistic of this test indicates the presence of interaction in the data.

A related technique has also been applied quite successfully in exploratory data analysis. One postulates that the model

$$y_{ij} = \mu + \alpha_i + \beta_i + k\alpha_i\beta_j + \epsilon_{ij} \qquad (41)$$

containing a multiplicative interaction term $\alpha_i\beta_j$ (or some multiple of it) fits the data. After determining that Model (35) is not adequate by examining residuals, we plot the computed residuals,

e_{ij}, from an additive model against the quantities $\hat{\alpha}_i\hat{\beta}_j/\hat{\mu}$, termed *comparison values*. This plot, called a diagnostic plot by Tukey, is examined for a linear pattern. If the residuals are linearly related to the comparison values by a line with slope b, a power transformation to the $(1 - b)$th power will alleviate nonadditivity. That is,

$$y_{ij}^{(1-b)} = \tilde{\mu} + \tilde{\alpha}_i + \tilde{\beta}_j + \tilde{\epsilon}_{ij} \tag{42}$$

will prove to be a good fit to the data, yielding small residuals, $\tilde{e}_{ij}$. The theory follows directly from Tukey (1949); a simple explanation is given by Snedecor and Cochran (1967, p. 331). Interactions are often difficult to comprehend and interpret; a transformation of the data that produces a simple additive fit can facilitate understanding.[17]

To continue our example, a diagnostic plot of the 13 pairs $(\hat{\alpha}_i\hat{\beta}_j/\hat{\mu}, e_{ij})$ is presented in Figure 17 and possesses no linear trend. A resistant fit to the points has this equation: Residual = 6.14 − 0.22 (comparison value). Since $p = (1 - b) = (1 - 0.22) = 0.78$ is nearly 1, we decide that no interaction is present.

There are many other EDA procedures for two-way tables that we shall not discuss. These include plus-one fits, direct addition of a multiplicative interaction term to Model (35), fitting other parameters, and the summarization of tables with ordinal factors by estimating linear fits for factor effects. We refer the interested reader to Tukey (1977) for an in-depth description of these alternative models and algorithms for their use in empirical situations.

DISCUSSION

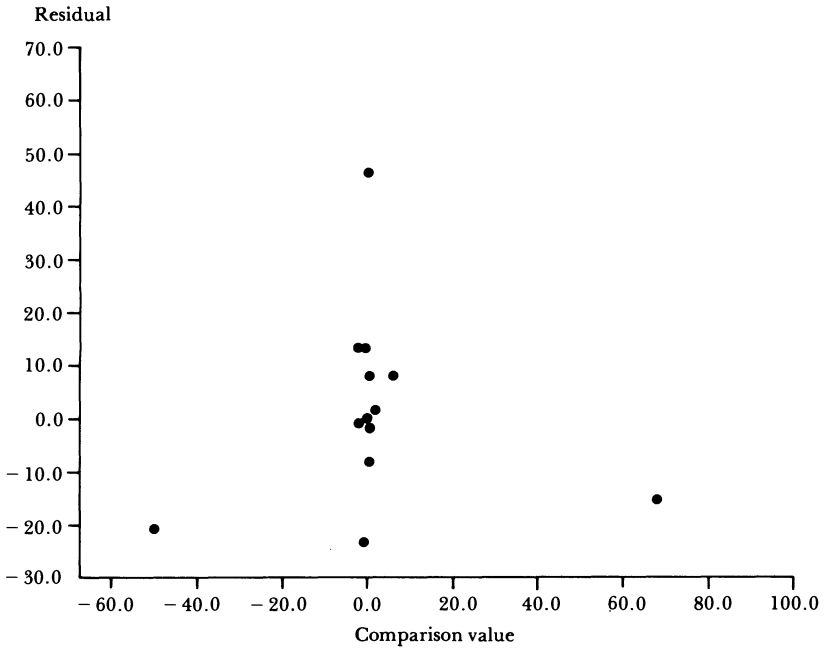
We have presented a small sample of procedures that John Tukey has created to facilitate and systematize the exploratory analysis of data, and we have tried to motivate their application when rationales were not self-evident or present in Tukey's original work. In a single chapter the constraints imposed by space

[17]An alternative exploratory procedure is to include the interaction directly as in (41). The procedure requires iteratively fitting

$$y_{ij} = \mu + \alpha_i + \beta_j + b(\alpha_i\beta_j/\mu)$$

to the data. Other models for two-way tables are described in Tukey (1977).

Figure 17. Diagnostic plot of median infant mortality rate data.



preclude a complete and detailed exposition of these and related procedures. In our selection we have aimed at presenting techniques that introduce topics likely to be useful to empirical sociologists, that provide a good feel for Tukey's general approach to data analysis, and for which there are commonly known parallel techniques in classic inferential statistics. But there can be no substitute for studying the source. While Tukey's text may prove difficult to some, the introduction provided in this review should help an interested reader penetrate Tukey's occasionally obscure style. Moreover, other writers are producing texts with simplified approaches and more traditional material. The following books contain examples of the application of EDA procedures: Erickson and Nosanchuk (1977), Fairley and Mosteller (1977), Hoaglin (in press), McNeil (1977), Mosteller and Tukey (1977), and Tukey (1977).[18]

[18]We have developed four modules for the Department of Housing and Urban Development as a curricular package. One module concerns EDA. See Leinhardt and Wasserman (1978) for details.

Exploratory data analysis has several important general features that make it particularly important to quantitative, empirical sociologists. One is that they are relatively resistant to the impact of poor-quality data. This does not mean that they will produce something from nothing, but it does mean that in the face of the problems typically encountered in empirical data, EDA procedures may help the analyst extract information from the data when traditional procedures would fail. In addition, EDA procedures help find transformations that are often essential if powerful classic statistical procedures are to be applied. Finally, the roles played by residuals and graphics in EDA present new ways of thinking about data that can lead to novel and more effective models and greater communication between the analyst and the analyst's audience.

The pedagogic utility of EDA should, by now, be apparent. The heavy use of graphics and simple iterative techniques permits students to *see* what is going on and get a feel for the process of analysis. The notion of pulling out partial fits provides a good intuitive grasp of complexity in functional forms of underlying relationships and, thus, fosters effective and creative analysis. The lack of reliance on abstract probability notions or complex mathematical forms provides for ready comprehensibility and mastery by students with minimal mathematical training. And, finally, the sequential development of ideas from the analysis of single batches, to multiple batches, to ordered multiple batches, to paired observational data, and to multiple carriers is logically consistent and leads naturally into a discussion of the model-fitting tasks of regression and analysis of variance. The application of this logical flow in an educational context should contribute to significantly improved quantitative training for students of sociology and to an increased proportion of students successfully mastering advanced quantitative methods.

REFERENCES

BARRODALE, I., AND ROBERTS, F. D. K.
1974 "Algorithm 478—Solution of an overdetermined system of equations in the l_1 norm." *Communications of the Association for Computing Machinery* 17:219–220.

BARTLETT, M. S.
1947 "The use of transformations." *Biometrics* 3:39–52.

BENIGER, J. R.
1978 "*Exploratory Data Analysis* by J. W. Tukey." *Contemporary Sociology* 7:64–65.

BISHOP, Y. M. M., FIENBERG, S., AND HOLLAND, P. W.
1975 *Discrete Multivariate Analysis*. Cambridge, Mass.: M.I.T. Press.

CRITTENDEN, A.
1975 "Vital dialogue is beginning between the rich and the poor." *New York Times*, 28 September, p. E-3.

CURTISS, J. H.
1943 "On transformations used in the analysis of variance." *Annals of Mathematical Statistics* 14:10.

DAVID, F. N.
1977 "*Exploratory Data Analysis* by J. W. Tukey." *Biometrics* 33:768.

ERICKSON, B. H., AND NOSANCHUK, T. A.
1977 *Understanding Data*. Toronto: McGraw-Hill Ryerson.

FAIRLEY, W. B., AND MOSTELLER, F.
1977 *Statistics and Public Policy*. Reading, Mass.: Addison-Wesley.

HINKLEY, D. V.
1975 "On power transformations to symmetry." *Biometrics* 62: 101–111.
1977 "On quick choice of power transformation." *Applied Statistics* 26:67–69.

HOAGLIN, D. C.
In press *A First Course in Data Analysis*. Reading, Mass.: Addison-Wesley.

HOAGLIN, D. C., AND WASSERMAN, S. S.
1975 "Automating stem-and-leaf displays." Working paper 109. Computer Research Center for Economics and Management Science, National Bureau of Economic Research, Cambridge, Mass.

KADANE, J. B.
1978 "Descriptive statistics." *Science* 20:195.

LEINHARDT, S., AND WASSERMAN, S. S.
1978 "Quantitative methods for public management." *Policy Analysis, 4.*
In press "Teaching regression: An exploratory approach." *The American Statistician.*

MCNEIL, D. R.
1977 *Interactive Data Analysis*. New York: Wiley.

MOSTELLER, F., AND TUKEY, J. W.
1977 *Data Analysis and Regression.* Reading, Mass.: Addison-Wesley.

SINGER, B.
1976 "Exploratory strategies and graphical displays." *Journal of Interdisciplinary History* 7:57–70.

SNEDECOR, G. W., AND COCHRAN, W. G.
1967 *Statistical Methods.* 6th ed. Ames: Iowa State University Press.

TUKEY, J. W.
1949 "One degree of freedom for non-additivity." *Biometrics* 5: 232–242.
1972 "Some graphic and semi-graphic displays." In T. A. Bancroft (Ed.), *Statistical Papers in Honor of George W. Snedecor.* Ames: Iowa State University Press.
1977 *Exploratory Data Analysis.* Reading, Mass.: Addison-Wesley.

WAINER, H.
1977 "John W. Tukey, *Exploratory Data Analysis.*" *Psychometrika* 42:635–638.

WELSCH, R. E.
1978 "Comment on Roberts." *The American Statistican* 32:52–53.