THE APPLICATION OF EXPLORATORY DATA ANALYSIS IN MARKETING: AN INTRODUCTION TO SELECTED METHODS

Armen Tashchian, Kennesaw College
J. Dennis White, Florida State University

Abstract

This paper introduces the family of techniques called exploratory data analysis. Unlike classical confirmatory statistics which rely upon strict distributional assumptions, parameter estimation, and hypothesis testing, EDA adopts an informal method of data examination designed to explore the structure of the data. Three representative EDA techniques are introduced and applications to marketing data sets are presented.

Introduction

Exploratory data analysis (EDA) is a generic term covering a large set of techniques which can help researcher develop a better understanding of the shape, form, and peculiarity of their data. The two major principles guiding EDA are the notions of openness and skepticism. Like a detective investigating a crime, EDA allows the researcher to be open to the unexpected and skeptical of the obvious and easy answer. The underlying premise is that the better you know your data, the stronger is your confidence in the appropriate choice of the analysis and the resulting conclusions.

Although the attitude underlying data exploration has long been an important part of the skilled data analyst's set of tools, the procedures have recently been popularized by John W. Tukey (Tukey 1977). Tukey's work has provided an impetus to numerous articles, books, and monographs on the topic (Cohen 1984; Hartwing & Dearing 1979; Leunhardt & Wasserman 1979; McGill, Tukey, & Larsen 1978), and several statistical packages incorporate the EDA techniques as a part of their standard offerings (Dixon & Brown 1981; Hogban & Peavy 1977; Nie and Hull 1981; Ryan, Joiner, & Ryan 1981; Velleman & Hoaglin 1981).

As the term implies, EDA is exploratory in nature. Unlike classical confirmatory statistics which rely upon strict distributional assumptions, parameter estimation, and hypothesis testing, EDA adopts an informal method of data examination designed to explore the structure of the data. Rather than relying upon numerical summaries, EDA uses visual displays to help identify structural characteristics of the data such as location, variation, skewness, bimodality, extreme values, and shape. Further, EDA employs robust and resistant statistics in identifying central and extreme values (Hoaglin, Mosteller, and Tukey 1983). It must be kept in mind, however, that EDA is designed to complement rather than replace the confirmatory data analysis. According to Tukey (1977), "Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone--as the first step" (p.3).

The purpose of this paper is to present the three most commonly used exploratory methods: the stem-and-leaf diagram, the five number summary, and the box plot. Examples from marketing and social science highlight the use and the detective power of these techniques. The methods presented here are a small subset of all the available EDA techniques. The interested reader should refer to Hoaglin, Mosteller, and Tukey (1983), Mosteller and Tukey (1977), and Tukey (1977) for complete coverage.

Displaying Data: The Stem-and-Leaf

In many cases, a data analyst may possess a data set in which one dependent variable is thought to be related to a variety of independent variables. If no theory is available as to the shape of the relationship among the variables, the analyst may need to examine visual plots of the data in an attempt to identify possible patterns of influence.

The stem-and-leaf display is a simple yet versatile method that enables the researcher to examine the data as a whole. It is easier to construct than its classical counterpart--the histogram, preserves the original data values, and takes a major step in sorting the data. The stem-and-leaf display can also be used to detect such features as: where the values are centered; the spread of the numbers; whether there is bimodality in the data; the number and behavior of values that are far removed from the rest; and, other data patterns that the researcher may not expect to see.

The basic construction of the stem-and-leaf begins by allocating a separate line in the display for each possible leading digit (the stem). The trailing digit (the leaf) of each data value is then entered on the line corresponding to the stem. For example, consider the following numbers which represent actual grades obtained by twenty two undergraduate students in a marketing class:

Grades = {78, 73, 78, 67, 70, 67, 62, 73, 75,
50, 78, 57, 93, 67, 80, 63, 60, 72,
77, 88, 94, 87}

The first digit of each number serves as the stem and the trailing digit (e.g., 0, 1, 3, etc.) functions as the leaf. The stem-and-leaf for the data is presented below.

5	07
6	772730
7	838035827
8	087
9	34

It should be obvious that each number can be recovered and that the data values are almost in order. The general rule for the upper limit of

the number of lines (categories) for a stem-and-leaf display can be approximated with the following formula:

$$L = [10 \log_{10} n] \qquad (1)$$

where n equals the number of data values and [x] is the largest integer not exceeding x. The above approximation was originally suggested by Dixon and Kronmal (1965) for establishing the numbers of categories for histograms. This rule gives reasonable estimates for the data values in the 20-300 range. For smaller samples (i.e., n<50), Velleman (1976) recommends the upper limit of number of categories to be computed using $L = 2\sqrt{n}$. The width of the stem-and-leaf display can be determined by dividing the range by the number of categories and round the quotient up the nearest power of 5 or 10.

Following these guidelines, the number of categories for the data set is approximately $L = 2\{\sqrt{(22)}\}=9$, the the width of the categories is [(94-50)/9]=5. The new stem-and-leaf display, incorporating these features is presented below.

2	5	07
5	6*	023
8	6	777
(4)	7*	0233
10	7	57888
5	8*	0
4	8	78
2	9*	34

The numbers in the far left hand corner indicate the frequency of observations in each class, while the value in parenthesis represents the frequency of the median class. The stem with a "*" represents the leaf values in the range of 0-4, and the stems without a "*" contain the leaf values in the range of 5-9.

Five Number Summaries

The best way to summarize a batch of numbers developed by the stem-and-leaf is to use the five number summaries: median, lower extreme, upper extreme, lower hinge, and upper hinge. These summaries are based on counts and as such, they are robust: an arbitrary change in a small part of the batch can have only a small effect on these summary values. Further, they indicate the amount of spread and help identify the outlying observations (See Table 1).

TABLE 1
FIVE NUMBER SUMMARIES

n - (number of observations)
M - depth of the median
H - depth of hinges
* - extremes

	median	
lower hinge	upper hinge	H-SPREAD
lower extreme	upper extreme	Range

Once the values in stem-and-leaf display are ordered, the researcher can easily find the extremes--the lowest and highest values (marked by "*"). The median (marked by "M") is the single middle value if the batch of numbers is odd and is the mean of the two middle values for even number of data points. The hinges (marked by "H"), the middle values between the extremes and the median, divide the batch into quartiles. The hinges and the median are resistant to the effect of stray values.

There are two convenient numbers that describe spread of the batch: range and h-spread. Range is defined as the difference of extremes, while h-spread is the difference between hinges (UH-LH). To define potential outliers, Tukey (1977) has suggested the following additional spread measures:

STEP = 1.5 * H-SPREAD, and
INNER FENCES = 1 STEP OUTSIDE HINGES

Values at each and closest to, but still inside, inner fences are labeled "adjacent". Values outside the fences cutoff points are generally regarded as outliers, deserving special attention. Using the examination data, the five number summaries are: M = 73, LH = 67, UH = 78, and H-SPREAD = 11. STEP is equal to 17, and the limit for inner fences if 50 and 95, respectively. In this illustration, the extreme values of 50 and 94 become the adjacent values. Table 2 displays the five number summaries for the total sample and male and female student subsamples.

TABLE 2
FIVE NUMBER SUMMARIES FOR CLASS EXAM EXAMPLE

n = 22	Total Sample			
M = 11.5		73		
H = 6.0	67		78	11
* = 1	50		94	44
n = 11	Male subsample			
M = 6.0		67		
H = 3.5	65		76.5	11.5
* = 1	57		94	37
n = 11	Female subsample			
M = 6.0		77		
H = 3.5	72.5		79	6.5
* = 1	50.0		93	43

The Boxplot

The purpose of boxplots is to summarize the information presented in the stem-and-leaf display and five number summaries. Boxplots show the structure of the batch in terms of location, spread, skewness, tail length, and outlying data. Further, boxplots permit simple comparison of several batches simultaneously.

To construct a boxplot, a rectangular box is drawn

with the ends defined at lower hinge and upper hinge and a solid line at the median. A dashed line (whisker) is drawn from each end of the box to the largest/smallest adjacent value in the batch. The outliers are presented individually by o's. Boxplots are usually displayed vertically, however, horizontal displays are common when several boxplots are being compared simultaneously. When several samples are compared, boxplots are usually drawn with varying widths to reflect the difference in number of cases for each subsample. McGill, Tukey, and Larsen (1978) suggest the width of the boxplot to be proportional to the square root of the respective sample sizes.

When comparing several boxplots, confidence intervals for the population median can be computed as follows:

$$m-1.58[H\text{-}SPREAD/\sqrt{(n)}]<M<m+1.58[H\text{-}SPREAD/\sqrt{(n)}]$$

where n denotes the number of observations in the boxplot. Two groups whose intervals do not overlap are significantly different at approximately .05. Since median, and hinges define the boxplot, it is resistant to arbitrary large data values. Actually, up to 25% of the data values can be made large without greatly disturbing the median.

Figure 1 presents the boxplot of the exam grades for the entire class as well as boxplots for exam grades broken down for female and male students. As can be seen from Figure 1, the boxplots show the location, spread, skewness, tail length, and outlying data values at a glance. The location is summarized by the median--the solid line inside the box. Spread is measured by the length of the box. The relative position of the median to the hinges is an indication of skewness. When the median is closer to the lower hinge, it is an indication of a positively skewed distribution. Similarly, when the median is close to the upper hinge it indicates a negatively skewed distribution. Tail length is indicated by the length of the "whiskers".

Examination of Figure 1 reveals that for the total sample, the distribution of exam grades is almost symmetrical. The scores for the female subgroup, however, have a smaller spread than the grades for the male students. While both distributions are skewed, the female sample is skewed to the left and the male sample is positively skewed. Finally, the two samples are significantly different from each other based on their median.

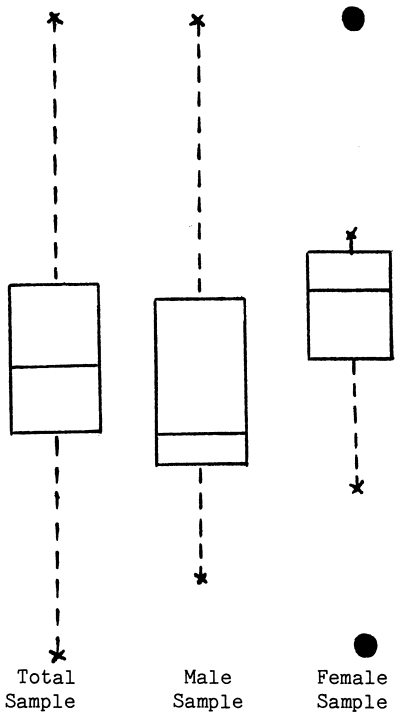
The next section presents two examples that show the application of the methods described above in actual research setting for studying symmetry of the data, skewness, bimodality and behavior of outlying values.

Illustrative Examples

Case-1

Diskin and Tashchian (1984) examines the impact of some selected demographic variables on tenant absorption of converted condominiums. Log-linear methodology was used to develop the final absorption model which included age, income and presence of children as three demographic variables. While fit of the overall model using chi-square statistics appeared acceptable, the confirmatory analysis did not hint if the assumptions of the model were violated. Thus, EDA techniques were used to study the distribution of standardized residuals in terms of normality and outlying values. The stem-and-leaf display and five number summaries are presented in Table 3.

FIGURE 1
BOXPLOTS FOR STUDENT GRADES



As can be seen from this table, the residuals closely follow a normal distribution. The median of the residuals is fixed at .05 and hinges are equally spaced around the median at -0.4 and 0.4. Further, all cases are within the inner fences of -1.6 and 1.6 and there are no outlying residuals.

TABLE 3.
RESIDUAL ANALYSIS FOR CONDOMINIUM EXAMPLE

	-1	33
	-0	88755
	-0	44443332211
	0	222223334444
	0	55668
	1	24

n = 36
M = 18.5
H = 9.5
* = 1

	0.05		
	-0.4	0.4	0.8
	-1.30	1.40	2.7

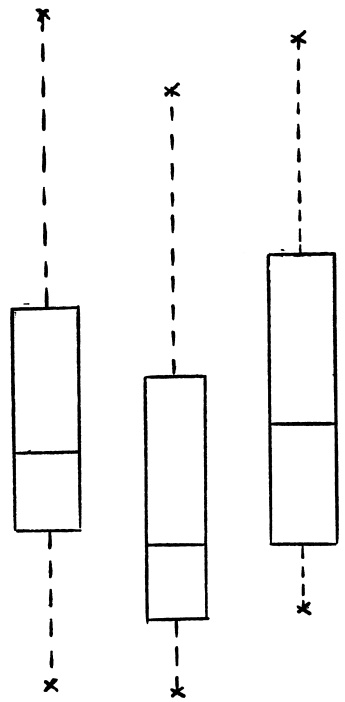
Inner Fences = (-1.6, 1.6)
Whiskers = (-1.3, 1.4)

Case-2

White and Carlston (1983) attempted to investigate the effects of memory accessibility of product information on consumer susceptibility to persuasion. The researchers predicted that susceptibility to persuasion would be dependent on the relative accessibility of supportive information in memory. Examination of the subjects' attitudes following receipt of the persuasive message, however, produced only mixed support for the original hypotheses. In an effort to discover what may have lead to the mixed support, an exploratory data analysis was performed. The results of the analysis is presented in Figure 2.

FIGURE 2
STEM-AND-LEAF DIAGRAM AND
BOXPLOTS FOR CONSUMER MEMORY EXAMPLE

1	-1	0
6	-0	00000
11	0	00000
35	1	00000000000000000000000
(10)	2	0000000000
27	3	000
24	4	0000000
17	5	00000
12	6	00000000
4	7	000
1	8	0



By examining the stem-and-leaf display, it is clear that the sample shows a clear bimodality, with the largest group of consumers displaying resistance to persuasion as predicted. The fact that the second grouping appears to be composed of a substantial number of subjects (as opposed to a few outliers), an attempt was made to discover the antecedents of the bimodality. Further analyses ultimately discovered that a consumer confidence variable accounted for the bimodal pattern.

Summary

This paper presented some selected techniques which facilitate the exploration of data. The reader is cautioned against applying these techniques with same rigid approach that has characterized much of traditional data analysis. Exploratory data analysis is interactive and researcher should be open to the unexpected and skeptical of the easy answer. When combined with classical, confirmatory statistics, these techniques can provide the marketing researcher with an excellent set of tools for discovering unique data patterns and trouble-shooting failed statistical tests.

References

Cohen, A. (1984), "Exploratory Data Analysis Methods: A Study of Industrial Worker's Work Role Centrality." Sociological Methods and Research, 12(4), 433-452.

Diskin, B.A., and A. Tashchian (1984), "Application of Logit Analysis to the Determination of Tenant Absorption in Condominium Conversion," Journal of the American Real Estate and Urban Economics Association, 12(2), 191-205.

Dixon, W.J., and R.A. Kronmal (1965), "The Choice of Origin and Scale for Graphs, " Journal of the Association for Computing Machinery, 12, 259-261.

Dixon, W.J., and M.B. Brown (1981), BMDP Biomedical Computer Programs, Berkeley: University of California Press.

Hartwing, F. and B.E. Dearing (1979), Exploratory Data Analysis, Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 16. Beverly Hills and London: Sage Publications.

Hoaglin, D.C., F. Mosteller, and J.W. Tukey (Eds.) (1983), Understanding Robust and Exploratory Data Analysis. John Wiley and Sons.

Hogben, D. and S.T. Peavy (1977), Omnitab II User's Reference Manual 1977 Supplement, NBSIR 77-1276, Superintendent of Documents, U.S. Government Printing Office, Washington,D.C. 20402.

Leunhardt, S. and S.S. Wasserman (1979), "Exploratory Data Analysis: An Introduction to Selected Methods," In K.F. Schuessler, (Ed.), Sociological Methodology, 311-365.

McGill, R., J.W. Tukey, and W.A. Larsen (1978), "Variations of Boxplots," American Statistician, 32, 12-16.

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

Nie, N.H., and C.H. Hull (1981), SPSS Update 7-9, New York: McGraw-Hill.

Ryan, T.A., Joiner, B.L., and Ryan, B.F. (1981), Minitab Reference Manual, University Park, PA: Minitab Project, The Pennsylvania State University.

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

Velleman, P.F. (1976), "Interactive Computing for Exploratory Data Analysis I: Display Algorithms," 1975 Proceedings of the Statistical Computing Section, Washington, DC: American Statistical Association.

Velleman, P.F., and Hoaglin, D.C. (1981), Applications, Basics, and Computing of Exploratory Data Analysis. Boston: Duxbury Press.

White, J.D., and D.E. Carlston (1983), "The Effects of Memory Accessibility on Resistance to Persuasion," Unpublished manuscript, Florida State University.