1-1-2012

Exploratory Data Analysis: A Primer for Undergraduates

Eric Waltenburg
Purdue University, ewaltenb@purdue.edu

William McLauchlan
Purdue University, mclauchl@purdue.edu

Exploratory Data Analysis:

A Primer for Undergraduates

by

Eric Waltenburg

and

William P. McLauchlan

Purdue University

N.B. This version of the manuscript will be supplemented with additional material later. Please do not quote or cite this without written permission from the authors. This version of the manuscript corrects several errors in the earlier 2010 version. This version of the manuscript replaces the 2010 version in its entirety.

Table of Contents

Errata

Although the citation was present, the 2010 version of this manuscript contained a formatting error that omitted the proper quotation marks around Professor Pollock's observation of the research process in chapter 2, p. 9. The 2012 version of the manuscript corrects this error.

The correct citation to Figure 2.1 (p. 25) is Pollock 2005, 9. The 2010 version of this manuscript omitted the page number. The 2012 version of the manuscript corrects this error.

The 2010 version of this manuscript did not note that the numeric summary described in chapter 3 (p. 32) was devised by Hartwig and Dearing (1979). The 2012 version of the manuscript corrects this error.

References in the 2010 version of this manuscript omitted Pollock 2009. The 2012 version of the manuscript corrects this error.

CHAPTER 1

AN INTRODUCTION TO EXPLORATORY DATA ANALYSIS

This book focuses on an approach and a set of methods for social scientists to use in the exploration of data. This approach or these methods are not related to any particular subject matter. However, the utility of the Exploratory Data Analysis (EDA) approach and the techniques presented here are widely useful for anyone who is approaching a set of data from the outset. The purpose of this is to provide a systematic scheme for looking at data and extracting the patterns that are contained in the data. To do that, we will outline a perspective or approach, a set of "steps," and a set of methods that can accommodate a wide variety of data. This will be an outline of a remarkable way of looking at social science data. The techniques are not new and they may seem rudimentary or basic, but they provide a unique set of techniques that precede much of the work that social scientists typically undertake.

Social science data are generally messy. The data are virtually never "normal." As a result, using traditional and rigorous statistical methods for assessing those data are not usually the best way to begin the analysis of data for a research project. This book is not a statistical treatment of methods for confirming or measuring relationships or testing hypotheses. Instead, this presentation should be considered a prequel that precedes any confirmatory statistical techniques one might wish to use on data. This examination begins with a philosophy of skepticism about data. The EDA approach assumes that understanding the data requires more than a cursory or superficial glance or mere collection of some quantitative data. It requires careful, systematic, and somewhat unique (uncommon) techniques.

The Philosophy and Process of Exploratory Data Analysis

Any treatment of Exploratory Data Analysis must begin with a clear statement about the philosophical approach to data. This might seem strange, but a questioning and open-mind is essential to using this kind of approach to the examination of data. Much research in social science fields is theory driven, with elaborate or strongly held

expectations about the relationship(s) under examination. Such an approach is designed to confirm or reject hypotheses about the relationship(s) under examination. This general approach to confirmatory research is well-developed in a great many social science fields. It has produced any number of confirmed or rejected hypotheses about relationships.

The use of exploratory techniques is intended to disclose patterns in sets of data. The purpose of these techniques is to uncover the shape and nature of the data you are examining. This perspective begins with "seeing" the data very closely or in detail, and examining it in broad or general terms as well. This approach involves no preconceived notions about any of the data, no matter how obvious the data seem to be on the surface. EDA expects one to find patterns in the data but that approach means, as John Tukey (1977: 1) indicated, engaging in "detective work." That is, for a comprehensive uncovering of patterns the focus should be on numeric, counted, or graphical detective work. (Tukey's book should be required reading for everyone interested in learning Exploratory Data Analysis techniques.)

The perspective of exploratory data analysis is described in a simple formula that Tukey (1977: 208) outlined:

Data = Smooth + Rough

A ***Smooth*** involves the relationship between two or more variables. The discussion about EDA techniques, however, begins with treatment or the examination of a single variable – a univariate analysis. The discussion in the next few chapters (before Chapter 5) will focus on the analysis of a single variable. The ***Smooth*** and ***Rough*** concept will become quite evident in the later chapters of this book.

This translates into the basic point that data should be considered in two parts. The first portion is called the "smooth" and that refers to the pattern(s) that can be extracted from the raw data using various techniques. EDA techniques focus on extracting the "smooth" from any set of data. The smooth, whatever it is, *comes from the data*, and is not derived from ~~our expectations or our guesses (hypotheses)~~ about the data. That means the first step in the EDA process is to extract the smooth from data. A single variable may have more than one pattern or smooth. Extracting the smooth from the raw data may require more than one pass through it and there may be more than a single pattern that the data contain. That is important to consider because it leads to a very basic point about EDA – the

objective of EDA is to **uncover the smooth in data**. Furthermore, always expect to find more in the data than first appears to be the case. Let the data speak for itself. Then, look again, at the rough to determine if there is a secondary pattern or smooth in the residuals.

The second portion of the formula, the ***Rough*** is the remaining residuals that contain no pattern at all. Residuals are what is left after all the patterns have been extracted from a data set. However, it is very important to look closely at the ***Rough***, because that set of values may well contain additional patterns that need to be examined. (That means one should **always analyze the residuals** that come from any analytic technique. There may be patterns in the residuals.) The techniques demonstrated later in this book should allow for the extraction of the smooth in data and any subsequent smooth found in residuals. That should leave the residuals with no pattern at all.

EDA begins by examining single variables. That is not the usual treatment that arises in connection with social science data. Multi-variate analysis often involves examining five, ten or more variables simultaneously. Here, we will devote a good deal of initial attention (Chapters 1 – 4) to exploring the shape of single variables in order to appreciate the nature of a variable and extracting as much information about the variable as is possible. That is the essential first step in EDA. **Get as much information as is possible from each variable**.

The core elements of EDA begin with a skeptical and blank sense of what the data may show. Even when you have some idea of what you might expect, you should look at the data carefully for pre-conceived notions may not be confirmed by the data. Always go through the data set, whether it involves a single variable or several variables, repeated times, until you are sure there in nothing more (no other smooths) to be found in the data. This requires very careful work and it requires attention to detail. The analyst needs to note even minor differences in data patterns. The result of this is a very sound and perhaps revealing understanding of the data that are being examined.

Much of this EDA approach is descriptive. This involves extracting patterns and the features of the data, rather than a summary or "what they add up to." Much of the EDA approach involves visual displays of the data that permit observers to understand the data better than other, statistical methods that indicate spread, distribution, and location in the data. These methods should be resistant to extremes or peculiar pieces of

the data. EDA methods should clearly indicate those outliers, but not be sensitive or pulled by those data points, **always pay attention to outliers**.

Displaying Data

One particular feature of this process should be emphasized at the outset. There is a core need to graph and display the data you are considering. A statement about a picture being worth a thousand words might seem trite, but there is no substitute for "seeing" the variable or the data. The graphical presentation of data is very important for both the analysis of the variables and for the presentation of the findings that emerge from the data. As a result, a good deal exploratory data analysis involves graphing and plotting data, both single variables and multiple-variable data sets. Tukey (1977: 126-31) presents a sophisticated and understandable discussion about (1) the selection of graph paper, (2) the decisions one should make regarding both the graphing (analysis) of the data, and (3) the need to display the graphs on tracing paper! Tukey's treatment indicates a now-dated perspective on the selection of graph paper and how to plot variables. That is because graphical software has proliferated over the years. However, the point he makes about "seeing" the variables and the data is still crucial (essential) because all software does not do an equally good job of plotting variables. Understanding how to graph data in order to analyze it and to present it clearly is still crucial to (1) understanding data and (2) EDA techniques. The discussion in Tukey and other sources (e.g., Cleveland, 1994; Robbins: 2005) indicates just how important it is to display raw data and transformed data carefully and accurately. More will be said about plots throughout this book, since the understanding of data and relationships often depends on the picture of the data.

EDA is focused on visual presentations rather than just numeric treatments. That is because "seeing" the patterns in the data is more likely to generate ideas or interest in exploration than having a numeric summary or distribution that is divorced from the data. As Hartwig & Dearing (1979: 15) indicate, visual displays of data emphasize the shape of the data and that is at least as important as the location and the spread. Visual presentations of the data exhibit the characteristics and the shape of the data much better than summary statistics. Lastly, any summaries of the data are dependent on the shape of

the data, and so the visual display of the data is the essential, first step in determining what summary (if any) to provide for a set of data.

The initial step requires examining (i.e., seeing) the patterns and shape of a single variable in an effort to uncover what the variable (the data) tells us about the variable it supposedly represents. Thus, whether the data show the time of day that a bus arrives at a particular stop on its route or the age of people residing in New Hampshire when they first voted in a primary election, the data should tell us something, hopefully a good deal, about the phenomenon they represent. That may sound simplistic, but as the discussion in subsequent chapters shows, sorting through a single variable and extracting the patterns the variable displays can be a lengthy but rewarding process. The variable may be apparently straightforward (i.e., already smooth). However, despite the usual assumption that data are normally distributed with reasonable and manageable spreads and dispersions, that smooth data rarely exists in the real world. In fact, the discussion in Chapter 2 illustrates the widely diverse kinds of data that most often face the social scientists who seek to understand and explain social and real-world relationships.

Plan for this book

The next chapter (2) focuses on the nature of social science data. Types of data will be defined, and explained as well as illustrated. That is an essential, first-step in understanding data. Always consider the kind of data you have, because that will determine the sort of analysis that can be performed and what you can "say" in the end about the patterns in the data. When data are collected, care must be taken to insure its accuracy and its nature.

Chapter 3 treats the kinds of analysis that should be performed on a single variable. This univariate (single variable) treatment is central to the first steps of EDA. This analysis of individual variables is important whether one has one or several variables that may be linked to one another. Those should be treated individually before they are combined in some way. The type of data involved in this analysis is important, and there are a variety of EDA techniques that can be applied to any type of data, with useful and interesting results.

It may be more interesting to explore the relationships between two or more variables, but first the nature of single variables is well worth exploring. That examination of a single variable will tell its own story. Chapter 4 will examine one variable in order to illustrate the nature of EDA techniques on a variable. This will show how the EDA techniques are used at the outset of any analysis – on a single variable. This discussion will also treat the variable – case filings for the federal district courts in 2005 – in order to begin the analysis or the exploration of these data in terms of patterns and treatments. This will show some patterns of the variable that warrant close attention and substantive explanation later. The value of this treatment of one variable is to illustrate some of the techniques of EDA analysis and to illustrate the generic social science approach to data analysis.

The next step in this exploration is to examine (see) the relationships that exist between variables. This will permit the extraction and analysis of ***Smooths*** and to permit both the analysis of ***Roughs*** and consideration of relationships between variables. Chapter 5 will provide a discussion of how plots and analysis of the relationships between two variables can be accomplished. Looking at one variable can be revealing, but the relationships between two variables can be quite different and interesting. If one is exploring such connections between data, then completing these kinds of analysis successfully are crucial to the determination of what kinds of relationships exist and the nature of those relationships. If the individual variable has been properly treated in step one (Chapter 4) then the analysis of two variables against one another may be straightforward. However, what this comparison reveals depends on what has been "done" to the individual variables and the patterns that have emerged from that univariate analysis.

Chapter 5 focuses on multi-variate exploratory analysis techniques. That is, this will examine a variety of methods that can be used when there are more than two variables involved in the analysis. Chapter 6 will explore the district court case filings data in light of various variables that might be of interest as explanatory factors for the patterns that emerge from the Chapter 4 analysis. In other words, Chapter 6 will apply the techniques presented in Chapter 5 to the data of interest in this exposition.

It should be obvious that this book will be useful to anyone who is really interested in understanding a set of data – one variable or more. Crunching numbers to get an "answer" is an approach used by many theory-driven researchers. There is nothing wrong with that. However, the first step, exploring the data, does precede the evaluation of any theory. Without such a exploratory analysis of the data, one cannot be confident that the patterns they have uncovered are really there. The basic proposition for this work is that we start with a variable or set of variables that we know little or nothing about. We do not hypothesize what the relationship between these might be. However, when we finish the exploration of the data we will know a very great deal about them and how they relate to one another.

CHAPTER 2

DATA AND VARIABLES

"Exploratory data analysis is detective work – numerical detective work – or counting detective work – or graphical detective work" (Tukey 1977, 1). Tukey presses his metaphor by pointing out that, like the detective investigating the scene of the crime, the social scientist needs both tools and understanding in order to mine the evidence and identify the salient clues leading to an explanation (solution). Much of the work of a social scientist does not involve anything like solving a crime, but rather the work involves exploring and understanding (explaining) a mystery. A detective might have little difficulty understanding that a crime has been committed – there is a body on the floor. The social scientist, however, must first specify the object of the investigation, then determine what the facts are and move on from there. Tukey's "social science detective" really must solve a mystery without the criminal analogy involved at all. Yet, the exploratory processes of the criminal and social science detectives are similar.

First, for both detectives, the facts surrounding the mystery must be identified. What happened, and what is the evidence? That is often a very important and almost the end result of the social scientist's explorations. "What happened?" is a question that deals with facts in a detective story and it deals with data in the case of the social scientist. Exploratory data analysis, therefore, is the first and essential step in the social scientist's investigation of a question of interest. Data can tell us a great deal about events and circumstances in the past and in the present. We may wish to predict the future, but no detective (or perhaps social scientist) would or could make very certain predictions about the future. So the job of the social science detective is to explore the past and the present as carefully and as systematically as possible.

That task – systematic and careful exploration of the past and present – comprises two essential elements. First, there is some *thing* about which you want to gain a deeper understanding (a phenomenon of interest). Second, there is evidence (data) associated with that thing to be collected and analyzed. In this chapter we will discuss how political scientists (really, social scientists generally) conceive of concepts in order to arrive at

systematic understandings of their phenomena of interest. This set of elements is much like seeking to determine just what the mystery involves and perhaps thinking about how to approach solving the mystery. You will be introduced to a variety of technical terms that are the foundation of political *science* research. A working knowledge of the points introduced in this chapter is essential to the exploratory data analysis enterprise.

From Phenomena of Interest to Variables

According to Philip Pollock, "the primary" goal of political science research is "to describe concepts and analyze the relationships between them." The challenge is to transform those concepts into concrete instruments so that they can be systematically analyzed (Pollock 2009, 8). Your phenomenon of interest is a concept. It cannot be systematically analyzed until it has been transformed into a concrete measurable instrument. This process is referred to as *operationalization*, and it involves four basic steps (see Figure 2.1; Pollock 2009, 8). First, you must determine your phenomenon of interest. What is it that you want to know more about? Now, taking this step might strike you as simple, but that is not the case. There is a difference between simple and *deceptively* simple, and identifying the phenomenon of interest is most assuredly the latter. Not just any phenomenon of interest will do. A suitable phenomenon of interest must be objective, empirical, and specific. Objective phenomena concern the verifiable, real state of the world or nature. They do not concern values or normative judgments as to whether that state is good or bad. Closely related to its objectivity, a suitable phenomenon of interest must be empirical. That is, it must be subject to observation. You must not rely merely on your faith, intuition, or common sense to determine the presence or magnitude of your phenomenon of interest. Finally, the phenomenon of interest must be specific enough that it permits systematic study. To simply state, "I want to know something about X" triggers the next obvious question: "What thing about X?" Only when you narrow and focus your phenomenon of interest can you begin to identify and collect salient data leading to a meaningful investigation of your mystery.

[Insert Figure 2.1 about here.]

Suppose we are interested in judicial process in the United States. That interest is very broad. However, such a general topic of interest is not manageable. We need to know more about the structure and the processes of the American judiciary before we can begin to examine a question that is of interest to us that relates to the U.S. judiciary. Perhaps we know enough about the general subject to know that there are 51 different judicial systems in this country, one in each state, and a federal system. Furthermore, perhaps we know that there are three levels of courts in the federal system – District Courts, Courts of Appeals, and the Supreme Court.[1] Let us begin to narrow the subject down and assume that we are interested in the U.S. District Courts and how they operate. The kinds of research questions (mysteries) we might explore about these courts are numerous and could occupy a lifetime of endeavor. We only have a semester; that means we need to focus more narrowly on particular questions relating to the District Courts.

The kinds of questions that could be asked about these courts (i.e., the mysteries) include how many judges sit on these courts, how they vote, what their backgrounds were (are), what kinds of decisions each of them reaches. There are certainly many questions about the kinds of cases these courts hear and decide, as well as how litigants fare (wins and losses) in these courts. Since there are 89 different District Courts organized into geographic units, there are all kinds of mysteries about differences among the districts in terms of their size, the judges, and the cases litigated before each one. Clearly, the social science detective needs to narrow the question (phenomenon of interest) down more than just the level of the court system that is to be examined. For illustrative purposes, let us focus on the workloads of the

A district covers either a state (such as Colorado) or a sub-state region (such as the Northern District of Indiana) Congress can reorganize the districts when it chooses to do that. Over time, a total of 89 districts have been created in the continental United States. This number does not include five additional district courts – in the District of Columbia, Guam, the North Mariana Islands, Puerto Rico, and the Virgin Islands. There used to be a federal District Court in the Panama Canal Zone, before the U.S. turned the Canal over to Panamanian control. In addition to the creation and elimination of districts, the Congress has complete control over how many district Court Judges are authorized to serve on each District Court

[1] One should realize that the other fifty court systems are organized in widely varying fashion, so that the Federal Court System is not typical of the "U.S. judiciary," which was our original, general focal point.

District Courts. More specifically, let us say our phenomenon of interest is the variation in each District Court's workload for 2005.

This is a very rough sense of how a general interest in a topic such as the "judiciary" can be focused and narrowed to the point where we have a mystery that can be detected and perhaps "solved." It requires a good deal of substantive knowledge about the court system in order to identify and focus the phenomenon that is of interest. That preparation is very important even though little time will be spent on those preliminary features of the research process here. However, this also requires a working knowledge of what data are available on District Court workloads.

Once the phenomenon of interest (the concept) is identified, the analyst moves to the second step in the process of operationalization; the phenomenon of interest is "nominally defined." Here, the empirical properties of the phenomenon of interest and the units to which the phenomenon applies (e.g., districts and years) are established. The workload of a court concerns the total number of actions, causes, suits, or controversies contested before a court.[2] Since our phenomenon of interest is the annual workload of the various district courts, our nominal definition of that concept is the total number of causes, suits, actions, or controversies (civil and criminal) contested in each district in a given year. Notice that in this definition, we have clarified the concept's empirical properties (all causes, suits, actions or controversies – regardless of substantive area), identified the units to which the concept applies (the 89 different District Courts), and specified the time period of interest (2005).

With the nominal definition in hand, the analyst begins to construct an "operational definition." That is, a strategy for the measurement of the concept is designed and an instrument is devised that measures the concept's empirical properties. Given that we have nominally defined a district court's annual workload as the total number of cases or controversies before the court for a given year, our measurement strategy could simply be to tally all the cases that are filed with each district in 2005. The instrument, then, would be the actual number of cases filed with each court for the specific year under analysis.

---

[2] *Black's Law Dictionary*, 6th edition.

Finally, empirical soundings are taken, and the operational definition is applied to the units being analyzed. This final step transforms the abstract concept into a concrete *variable* with empirical properties that vary from unit to unit. So, for example, in 2005, 669 cases were filed in the U.S. District Court of Maine; 12,545 cases were filed in the Southern District of New York; 2297 cases were filed in Northern District of Indiana.

Reliability and Validity

How a concept is operationalized has a tremendous bearing on the results of whatever type of analysis is being performed. Simply put, if your operationalization is faulty, your conclusions will be faulty. As in computer programming, GIGO is at work in exploratory data analysis. Your operationalization should produce instruments for exploring the mystery you have identified that are *reliable* and *valid*. Reliable instruments are the result of a measurement strategy that yields the same sounding of the concept for a given unit every time the concept is measured for that unit. A valid instrument is one that actually taps the concept it is supposed to measure. Obviously, unreliable and invalid instruments will produce results that are unpredictable and erroneous.

"GIGO" refers to "Garbage In, Garbage Out."

As a silly, but hopefully useful, example, consider a set of directions given to 20 different partygoers that must be followed in a specific order to arrive at the desired destination. If each set of directions was randomly ordered, the partygoers might well arrive at 20 different locations. Clearly, this is an unreliable set of directions. Alternatively, consider the same 20 partygoers given 20 identically ordered sets of directions; in each case, however, south is mislabeled as north. Here, all 20 partiers would arrive at the same, albeit wrong, location. If the set of directions is an instrument measuring the concept of the proper route to the party, this is an invalid measure.

As the preceding example suggests, it might seem conceivable to have unreliable instruments that are valid or reliable instruments that are invalid. According to Kellstadt and Whitten, however, reliable but invalid instruments and vice-versa are, for the purposes of systematic, scientific knowledge, not possible (2009, 95-96). This is because of instrumentation's crucial role in the scientific process. Recall that your phenomenon of

interest began as an abstract concept that had to be transformed into a concrete measurable variable in order for it to be systematically analyzed and understood. If the measurement of this concept is unreliable, any understanding is brought into question, regardless of how valid the measurement is. After all, the variable's value for any given case cannot necessarily be reproduced. Our understanding of the given phenomenon of interest is a prisoner of our measure of that concept, and that measure can change haphazardly. Similarly, invalid instruments do not measure what they purport to measure. Consequently, any understanding born of these instruments is dubious. They can be utterly reliable, but in the end we acquire no meaningful understanding of our phenomenon of interest. Ultimately, both reliability and validity are necessary conditions for acquiring a systematic understanding of the phenomenon of interest. Neither, however, is sufficient.[3]

In terms of our district court example, the number of square feet in the Courtrooms for each district would be a reliable measure. A square foot in the Northern District of Illinois is identical to a square foot in a courtroom in the Eastern District of Washington; however, square footage would certainly NOT be a valid measure of court workload. The number of cases filed in each district in a year, is both reliable and valid, as an indicator of workload. As a count, each filed case is equivalent in all districts. This means that it is a reliable indicator of district court workload. It is also valid, as an indicator of workload, since each district must treat each case the same.[4]

Units of Analysis and Levels of Measurement

Two more foundational bricks concerning measurement must be put in place before we can begin to build a systematic understanding of the phenomenon of interest –

[3] There are tests that can be performed to assess a measure's reliability and validity. The interested student can consult Johnson, Reynolds, and Mycoff 2008, pp. 94-104; Salkind 2006, pp. 105-118; Nachmias-Frankfort and Nachmias 2007, pp. 148-157.

[4] Each case, however, does not actually represent the same amount of work for judges. Simple cases can be resolved more quickly and more easily than complex cases involving a multiplicity of litigants. The parties settle some cases after the filing but before going to trial or before requiring any attention by a judge. So there is a great variation in the amount of work required for each case that is filed in district courts. For purposes of this study however, case filings will be used to indicate workload.

units of analysis and levels of measurement. A *unit of analysis* is "the entity (person, city, country . . .) we want to describe and analyze; it is the entity to which the concept applies" (Pollock 2009, 52). Another way of thinking about the unit of analysis is that it is the entity that is being measured in the operationalization process. For example, if your concept is the understanding of the information presented in a given section of the course for a particular class of students, and you choose to operationalize that concept using performance on an exam, your unit of analysis would be each individual member of the class. Returning to our running example of District Courts' workloads, since your phenomenon of interest is the workload of these courts, your unit of analysis is the individual district court. That is, you would measure the caseload of each of the 89 districts. Or, to put it even more directly, you would determine the caseload of the Maine District and the Massachusetts District. So on through the last district. Since your unit of analysis is the court, and there are 89 of those, the value of the variable "caseload" would be recorded 89 different times – a value for each district.

Some studies may employ a unit of analysis that varies over both space and time. Studies of this sort are called cross-sectional time series studies. For example, your phenomenon of interest might be the annual caseload for each federal district court for the years 1990 through 2000. In this example, the unit of analysis varies over space – each of the 89 district courts – and time – each year between 1990 and 2000In the court example, the unit of analysis varies over space – here, quite literally geographic space.

The student exam example also varies over "space" in the sense that each student is a different physical specimen.

Each of the 89 districts is tied to a unique geographic area, and the phenomenon of interest is being measured across each of those geographic areas. Analyses of this sort are called *cross-sectional* analyses. It is possible, however, that the phenomenon of interest contains a time dimension. That is, the interesting variation in the concept occurs over time. These studies are called *time-series* analyses. In time-series studies, the unit of analysis would be some discrete temporal period (e.g., months, years, decades), while remaining constant with respect to space. For example, the phenomenon of interest might be the monthly caseload of the Minnesota District Court in 2009. In this instance, the caseload variable would be measured for that one district for each month in 2009. That is, the value of the caseload variable would be recorded 12 different times, once for each month in 2009. (See Kellstadt and Whitten 2009, 23-26 for a discussion of space and time dimensions as they pertain to the unit of analysis.)

Some studies may employ a unit of analysis that varies over both space and time. Studies of this sort are called cross-sectional time series studies. For example, your phenomenon of interest might be the annual caseload for each federal district court for the years 1990 through 2000. In this example, the unit of analysis varies over space – each of the 89 district courts – and time – each year between 1990 and 2000.

A variable's *level of measurement* identifies the nature of the information contained in the operationalization of the concept. There are three levels of measurement – nominal, ordinal, and ratio – and each imparts a different amount of information. *Nominal* levels of measurement simply classify or categorize the cases. Numerically, the values associated with each case – the values over which the variable varies – are meaningful only in the sense that the values associated with different cases differ from one another. The magnitudes of the different values mean nothing. Although one case may have a greater numerical value than another case, that does not mean there is a ranking or ordering of the cases regarding the phenomenon being measured. Rather, the values simply identify different qualities of the phenomenon. The only requirement of the values associated with the different qualities of the phenomenon in a nominal level of measure is that the values be mutually exclusive and that the different categories of the phenomenon being represented numerically are exhaustive.

Many research design texts include "interval" as a fourth level of measurement. Interval levels of measure possess all the properties of ratio levels of measurement except for a true, meaningful zero value. In political science there are very few concepts that can be operationalized at an interval level but not at the ratio level. Accordingly, we will dispense with a detailed discussion of this level of measurement. Interested students should consult Salkind 2006, pp. 100-105; Trochim 2001, pp. 103-105.

An example of a nominal level of measure is a variable that would measure the identity of each district court. In this case each district would be assigned a unique value, 1 through 89. Other than identifying the different districts, the values are meaningless. For example, the Maine District Court could be assigned the value "1" however, it could also be assigned the value "43" or "89." That would separate it from all the other districts, but the value would not measure quantity or magnitude. These "numbers" only represent different parts of the county.

Ordinal levels of measure add ordering or ranking information for the cases with respect to the phenomenon of interest. That is, like the nominal level of measure, ordinal instruments classify and categorize, but they also convey information that permits the analyst to identify cases that are greater than or lesser than other cases with respect to the concept they measure. As such, ordinal levels of measurement require that the values associated with different cases reflect the ranking of the concept for that case. That is, cases that have "more" of the phenomenon being measured should be assigned a value that is greater than the value assigned to cases with "less" of that phenomenon. The specific width of the intervals separating those values, however, is meaningless. As a consequence, although the values 1, 2, 3, and 4 convey a numeric ranking, the analyst cannot say that cases assigned a value of 4 have twice as much of the property being measured as those cases assigned a value of 2, only that the case assigned a value of 4 has more of the property than the case assigned a value of 2. In the case of square footage, a rank ordering of all 89 districts based on their total courtroom square footage where districts with more square footage are assigned higher values than districts with less square footage would be an ordinal indicator of the district with the most and the least courtroom area. However, that rank order indicates nothing more. That is, the district that is assigned the value of 20 does not have five times the square footage as the district assigned the value of 4.

Numbers are typically used with nominal level variables. However, letters could also be used or even symbols. All that is required is that each category present in the variable be represented with a different number, string of letters, or symbol.

The unequal intervals between the values assigned to the "low," "medium," and "high" response is intentional and intended to demonstrate that ordinal measures require only that the numbers associated with the categories reflect the order of the categories. The magnitude of the difference between the categories is irrelevant.

A better example of an ordinal level of measure might be derived by surveying district court practitioners regarding their impression of the amount of cases contested in each court in a "typical" year. The practitioners' possible responses would be "low," "medium," and "high," to which you could apply the values 1, 7, and 26.

Finally, *ratio* levels of measure convey the greatest amount of information about the concept being operationalized. The values associated with a case have the full

mathematical properties of numbers. That is, they categorize, order, and permit ratio comparisons. A case assigned a value of 10 is in fact twice as great with respect to the phenomenon being measured as a case with a value of 5. This type of comparison is possible because this level of operationalization, at least theoretically, has a true zero point – the complete absence of the phenomenon being measured.

As an example, again consider the original operationalization of our caseload variable. If the value of this variable is the actual number of cases on each district court's docket, then this operationalization is a ratio level of measurement. Using this operationalization, say the Nevada district court has a caseload value of 500, while both the Oregon and Colorado districts have a caseload value of 1000. In this instance the analyst could say the workloads of the Oregon and Colorado districts are twice the size of the Nevada District's workload. Such a comparison can be drawn because, at least theoretically, it is possible for a district court to have zero cases on its docket.

Ultimately, one should strive to operationalize concepts at the highest level of measurement that is possible. Higher levels of measure convey more information about the concept. If a variable is operationalized at a higher level of measure it is always possible to "collapse" the variable down in terms of its level of measurement. After all, ratio levels of measure contain all the information contained in ordinal and nominal level operationalizations. It is not possible, however, to do the opposite. Nominal levels of measure, for example, do not contain the information necessary to order the cases or to make ratio comparisons. Finally, more *analytic* techniques are available to treat concepts operationalized at higher levels of measurement. This point will be emphasized and clarified in subsequent chapters.

Some concepts can only be operationalized at certain levels. For example, the concepts of sex and race can only be categorized. Values can be applied to the different categories (0 = male; 1 = female; 0 = White; 1 = Black; 2 = Hispanic; 3 = other), but the magnitudes of the values are meaningless. Despite having a value of 1, women are not greater than men with respect to sex.

Data Collection

The purpose of this study is not to have students collect data, but to understand its analysis. However, a few words should be said about data collection. The source of one's data may be an Internet site, a volume in a library, or a set of in-person interviews that are converted into quantitative data – that is, systematically examined so that concepts are extracted and measured (operationalized). One should recognize that the data must be managed and retained to insure two general goals. The first goal is to insure that the researcher can explain and understand (remember) exactly what the data mean, where they came from and what decisions were made with respect to operationalization and data processing and manipulation. That should permit the researcher to return to the data some time (years?) later and conduct additional research using the data or add additional data to the data set and complete an extended research project using the supplemented data. Second, any other researcher should be able to replicate the first study, using the same data.

The empirical research process in the social sciences, not unlike any research process, requires the researcher to be systematic and deliberate throughout the steps in the process. The researcher must begin with a clearly articulated research question of interest and a set of specified variables that are expected to shed light on the question. The researcher must identify sources for the variables. These sources could be personal interviews with selected respondents. The sources may be library materials or archival documents. The source of data that capture the essence of variables could be already quantified material published in the public domain or under copyright. Some of this latter category might well be drawn from already existing data sets that other social scientists have created and made public, or these could be available on the Internet from various external, even invisible sources. The most careful consideration about developing sources of data is to determine how accurate and reliable the data are. It the data are drawn from someone else's work then the prior work has to be evaluated very carefully. If the researcher is creating their own data then, the creation of the data must be done very carefully.

For the following discussion, let's assume one is collecting one's own data. Collecting quantitative data involves a set of steps, all of which need to be documented carefully. A primary reason for the documentation of the collection process is that other

researchers may wish to use these data. In order for these researchers to have confidence in the data's reliability, they need to understand how the data were collected. In addition, the original researcher may wish to revisit the data at a later time (perhaps years later) in order to reanalyze the data, add additional variables to it, or up-date the data so that additional cases can be included in the analysis. It is important to remember that a scientific process involves the ability to replicate a study or re-do the study and presumably obtain the same results. That requires that every step and every decision in the data collection process must be documented.

Unfortunately, researchers often do not follow this level of detail, and this can prove fatal at later stages in the research process. So, keep very careful track of (1) the sources of the data, (2) how the data were processed, and (3) what each data value means or what it stands for. Consider the District Court workload example we have employed throughout this chapter. To operationalize workload, we chose to use the total number of cases filed with each District Court in 2005, and consequently, we need to identify the source of these data. Annually, the Administrative Office of the United States Courts publishes *Federal Court Management Statistics*. This publication contains a variety of data for each district court, including the total incidence of case filings.

Constructing the data in a form that can be analyzed involves putting the data into a "machine-readable" form usually by entering it onto some kind of "spreadsheet" on a computer. Before one does that, however, one should construct a code book that identifies each variable and what each value of each variable "means." For the district court data we have discussed throughout this chapter, the codebook might look fairly simple since few variables will be in the data set. The codebook might look something like this for one of the variables:

"Annually" does not necessarily mean the same twelve-month period for all purposes. There is the calendar year, beginning in January and ending at the end of December each year. There are Fiscal Years, that can be any twelve-month period, and for the federal government that annual period starts on October 1 of each year and ends September 30. For many states and other organizations the fiscal year begins July 1 and ends June 30 of each calendar year. The statistical reporting year for the data source discussed here is October 1.
One needs to be careful if the data collected, from different sources, involves annual data that do not correspond to the same 12 month period.

districtcourt = the identity of each u.s. district court

1. AK = Alaska District Court
2. ALMD = Alabama Middle District
3. ALND = Alabama Northern District
4. ALSD = Alabama Southern District

So forth

In the above example, "districtcourt" is the name of the variable. Notice that a brief description of the variable is attached. This is referred to as the "variable label," and it provides information as to what concept the variable is intended to measure. When possible, descriptive variable names should be used. Some concepts, however, are too complex to be pithily identified, and in these cases the codebook and variable label are especially useful. In addition to the name of the variable and variable label, the codebook entry should include the values of the variable as well as what those values represent. In this example, districtcourt would have 89 different values, and each value represents a specific district. The name of each district is a "value label."

All the variables in the data set should be present in the Code Book, and any coding decisions about a variable should also be reported in the Code Book. Then when the data are entered in the spreadsheet the researcher or anyone else can determine what each cell contains.

As noted above, data are entered in the form of a spreadsheet composed of columns and rows. Each column identifies a variable. Each row identifies a case (unit). Each cell (the point at which the column and row intersect) is a particular case's value for a given variable. When entering data in an actual spreadsheet program such as Excel, always put the labels for all the variables in the data set in columns across the first row of the sheet. It is also extremely useful to use the first column in the spreadsheet to identify each unique case. This helps to ensure that the correct values for the variables are entered for each unique case.

Accordingly, in our example, the first few columns and rows of our data set would look something like what is displayed in Table 2.1.

[Insert Table 2.1 about here]

The data are displayed in a spreadsheet, and they should be collected using a spreadsheet that is easily accessible and understandable by other researchers.

If one wanted to add additional variables to this data set, one could do it in the columns to the right of CASE FILINGS. Thus, if one wanted to add the number of judges in each district and the Circuit each district was in, that would be relatively easily done. Simply extend the data set out two more columns, one for the number of judges and one for the identity of the circuit. The cases stay the same. Thus, the first case in the data set was and is the Massachusetts District Court. The cell that is at the intersection of the second row/fourth column contains the value of the number of judges sitting on the bench of the Massachusetts District Court. Similarly, adding more data is also relatively easy if the data are drawn from the same data source. [5] Say we wanted to add another year of workload to our data set. This would be accomplished by appending those data to the bottom of the spreadsheet. In this instance, the identity of the variables would not be changed. There would simply be more rows (units or cases) for which we had data.

It is absolutely essential that a researcher maintain a codebook. That codebook must contain several pieces of information. First, a list and definition of each variable should be in the codebook. In addition, the values for each variable should be defined so that a subsequent data user or collector can tell what a variable's value means. Thus, for example, NYED needs to be defined in the codebook. If each District court is given a separate "number" such as "1" or "89" those numbers also need to be identified or defined in the codebook. Any coding decisions that the collector makes during data collection need to also be recorded in the codebook. The importance of this cannot be stressed enough. The value of a codebook is that the researcher will have a complete record of the data set and its collection. That means the researcher can return to the data

---

[5] If you wish to add additional cases from a different data source, you must be very careful to insure that the data source provides the same data, for the new YEAR. There are other possible sources of these data, but they may not "count" Case Filings" the same as the Federal Court Management Statistics does. The District Court, the Number of Judges and the Circuit variables could be easily obtained from other sources and would be identical to these data in the Management Statistics source.

It is interesting to note that the Circuit designation for some districts changed in the early 1980s. That was when the 11th Circuit was created, and the nine district courts in Alabama, Florida, and Georgia were moved from the 5th Circuit to the 11th. So if you were adding data from 1975, the Circuit designation for these courts could be the 5th Circuit rather than the 11th.

Incidentally, the purpose for adding the "Judges" and the "Circuit" would be because the researcher wanted to examine the Case Filings categories by the Circuits in which these data arose. The Circuits, although designated by number, are only a categorical variable. Perhaps controlling for the number of judges in a District would also provide a "better" indicator of the amount of work each district was faced with. So "Judges" might be considered a control variable, rather than an independent variable.

later (months or years later) and be able to understand the data and its collection. In addition, another researcher could reconstruct the data collection and analysis process if the codebook is complete. Furthermore, additional variables could be added to the data set later, if the codebook is complete and exhaustive.

At the conclusion of data entry, the researcher needs to insure that the data are correct. Entering quantitative values for variables in a spreadsheet is bound to involve some typographical errors. Not checking the data at the end of the data entry means that later, when those inevitable errors are uncovered, the researcher will have to return, virtually to the beginning in order to conduct all the analysis and research over again, using the corrected data. Doing some preliminary analysis to insure that you have all the districts (all 89) and that the Case Filings variable does not contain one or more entries that are completely out of line, such as a "1" or a "15008997." Check for missing data or empty cells. Filling those in at this point is much easier than doing it later.[6] A researcher must "look" at the data that were collected. That means visually inspect the spreadsheet and not obvious errors.

Summary and Conclusion

The social science detective employing exploratory data analysis attempts to mine as much information from the data as possible, free from the constraints of preconceived expectations and theories. That is, the detective lets the data speak for themselves. Before the data can speak, however, the detective must give them a voice. First, the social scientist must specify the phenomenon of interest. Once the phenomenon of interest is specified, the social scientist *operationalizes* it so that its empirical properties are measured and it can be analyzed. In the process, salient data are collected, and the abstract phenomenon of interest becomes a concrete *variable*.

Hartwig and Dearing define a variable as "a set of values each of which represents the observed value for the same characteristic for one of the cases being used in the

---

[6] This stage in the process can be quite tedious. It is like dotting the "i's" and crossing the "t's." but sloppy work will yield GIGO. The conclusions derived from "garbage" are probably completely wrong or at least subtly misleading. Not correcting obvious or non-obvious mistakes in data collection is no different than making up the data altogether and that reprehensible behavior is not worthy of comment.

research" (1979, 13). In other words, a variable comprises the range of values of an operationalized concept. Each individual case under investigation (i.e., each entity to which the concept applies) will have a measured value for that variable. Take for example, the workload variable we have discussed throughout this chapter. Here, the unit of analysis is the individual district court (n = 89), and the variable (*caseload*) has a set of observed values ranging from 485 (the observed value for the district court of Alaska) to 17099 (the observed value for the Eastern District of Pennsylvania). Moreover, since *caseload* is measured at the ratio level, we can say that the Eastern District of Pennsylvania's workload is more than 35 times greater than the workload of the Alaska district court.

The social scientist employing exploratory data analysis takes these and other facts pertaining to variation in the district courts' caseloads and asks what is happening here? The point of departure to answering that question is a detailed examination of the single variable, *caseload*. Identifying the shape of the data with respect to that single variable, describing its distribution, and understanding its nature can reveal a great deal about the phenomenon it represents – the magnitude of each district court's workload. In the next chapter, we discuss a number of techniques to display, summarize, and understand the distribution of data on a single variable.

Table 2.1 Sample of Data Collected and Arrayed on a Spreadsheet.

districtID	districtcourt	caseload
MA	1	3633
ME	2	669
NH	3	692
RI	4	705
CT	5	2424
NYED	6	7136
NYND	7	2001

Figure 2.1 The Preliminary Stages of Developing a Research Project.

Framework	Question of Interest	Phenomenon of Interest	Operationalization	Variable
Empirical Project Statements	What is the Workload of the U.S. Courts?	Caseload Patterns in U.S. Federal Courts	Number of Filings in Courts will indicate their workload	The number of new cases filed in the Federal District Courts in 2005
Normative or Prescriptive Statement	The Courts are overworked and do not provide justice to litigants.	*		

* The normative statement about "justice" and the Courts is not amenable to empirical or objective research. No empirical research can provide any sort of answer for a question like this because the "answer" depends on one's preferences or definition of "justice." Even the term: "overworked" is a normative statement and cannot be measured objectively.

Source. Adapted from Pollock (2005, 9).

CHAPTER 3

Univariate Exploratory Data Analysis

In its most basic sense, exploratory data analysis is concerned with the identification or discovery of patterns pertaining to a variable. Recall, a variable comprises the range of values of an operationalized concept for the cases under analysis. *When those values are arranged in numerical order, they form a "distribution"* (Hartwig and Dearing 1979, 13). Exploratory data analysis, then, involves the systematic search for patterns in each variable's distribution. This search for patterns is the first step towards answering the question "What happened here?"

The systematic search for patterns begins with an intimate understanding of a variable's distribution. In other words, the analyst seeks to determine what the data underlying a variable look like. What is a variable's most common or typical value? How different are each value of the cases under analysis from one another and from the typical value? Are there cases that have extreme values on the variable? Are the cases symmetrically (evenly) distributed around the variable's most common or typical value, or do cases "pile up" at one end of the distribution or another?

These questions focus on three fundamental characteristics of a distribution – its measures of central tendency, its measures of dispersion, and its shape. In this chapter we will examine those characteristics in some detail. We will look at the various ways in which these characteristics can be indicated or assessed. There are various measures of central tendency and dispersion that represent a variable's characteristics. The ways these indicators are derived and what they mean about the variable are very important questions for uncovering the "mystery" in the data. We will identify the strengths and weaknesses that are associated with different summary statistics, and we will discuss the relationship between a distribution's shape and the validity of different measures of central tendency and dispersion. In the process we will make the case for utilizing a graphical depiction of a distribution in order to develop a more complete understanding

Hartwig and Dearing likewise refer to three characteristics of distributions. They identify them as the **location**, **spread**, and **shape** (1979, 13).

of the data underlying a variable. Each of these points, and the related techniques, are then illustrated on the federal district court workload data we introduced in Chapter 2.

Measures of Central Tendency

Univariate summary or descriptive statistics are just what their name implies – a single number that represents an important feature or characteristic of one variable. The *mode*, the *median*, and the *mean* are summary statistics that report a variable's most typical value, or as Hartwig and Dearing put it, "the point at which the distribution is anchored, or located" (1979, 13). To one degree or another, each of these summary statistics reports a value that is representative of all the values in a variable's distribution. The *mode* accomplishes this by reporting the value in a distribution that occurs with the greatest frequency. In a statistical sense, the mode is a distribution's most probable value; that is, the value that is most likely to occur. As an example, consider placing 100 poker chips in a hat. Seventy-five of the chips are red; 15 of the chips are blue, and 10 chips are green. Given this distribution of poker chips, red is the modal category, and if one were to blindly draw a chip from the hat, red would be the most likely color selected. The value of the mode is determined by simple visual inspection. There is no computational formula employed to derive it. Finally, the mode is a summary statistic appropriate for all levels of measure; it is the only summary statistic for variables measured at the nominal (categorical) level.

The *median* is a positional measure. It is the value or potential value of the case in a distribution above and below which exactly 50 percent of the distribution falls. In other words, the median is the 50^{th} percentile. It is the balancing point in a distribution. Thus, in a distribution with an odd number of cases, it is the value of the middle case. In a distribution with an even number of cases, it is the value midway between the values of the two middle cases. To illustrate, consider the distributions depicted Table 3.1 below.

[Insert Table 3.1 about here.]

Distribution 3.1A has an odd number of cases, seven to be precise. Therefore, its median is the value of the middle case – here, the fourth case in the distribution. There are as many cases positioned above it as below it. The value of this middle case is 23, and we

highlight it in yellow to better illustrate the median. Distribution 3.1B has an even number of cases, specifically six. Consequently, its median falls at the midway point between the values of the two middle cases – i.e., the third and fourth cases. Since the values of the middle cases in Distribution 3.1B are 22 and 23 (highlighted in yellow), the value of the median is 22.5. This value is derived by adding together the values of the middle cases and dividing by two. Because the median is a positional measure, it requires variables that can be ordered. The variable must be operationalized at the ordinal level or above. Finally, and perhaps of greatest significance, the median is a "resistant" summary statistic. That is, extreme scores in the distribution do not affect its value. As a result, the median is the most representative value of a distribution in the presence of extreme scores, or "*outliers*."

An outlier is commonly defined as a case whose value on a variable falls well beyond (either above or below) the typical pattern of the other values on that variable. In other words, it is a case whose value results in it standing at some distance from the other cases in the distribution. Generally, there are three explanations for the appearance of an outlier. First, the extreme score producing the outlying case might be the result of a simple processing error. Perhaps during data collection a value was recorded incorrectly, or perhaps during data input a value was entered incorrectly. For example, you may have intended to enter a score of 10 and instead struck the 0 key several times too often, entering a score of 10000 instead. A second explanation for outliers is that they are the result of your operationalization. Perhaps the way you chose to measure a variable results in a case or a handful of cases taking on extremely high or low scores. Finally, the outlying case or cases might genuinely reflect the data's distribution. In other words, extreme scores are truly present in the data. The distribution of personal income in the United States, for example, does include the cases of Bill Gates and Warren Buffet. These are extreme, but valid outliers.

Using summary measures to describe a distribution in the presence of outliers will result in wholly wrong understandings of a distribution's nature. Consequently, distributions should be carefully examined for the presence of outliers, and when they are found, their cause should be ferreted out. If the outlier is a simple processing error, fix it. If the outlier is the result of your operationalization strategy, you may attempt to devise a

different operational definition. Of course, an alternative operationalization may not be possible. If that is the case, then your choice is to acknowledge the outlier and conduct your analysis in full recognition of its presence. To do so, you might exclude the outlying case(s) or retain the case(s) and treat them differently or explain them separately from the bulk of the data. This might call for the use of more resistant summary measures and analytic techniques. These choices are true as well when the outlying case is genuinely present in your data. Simply put, the data cannot be "fixed," and your operationalization should not be altered. Thus, your choices are either to drop the case(s) or retain the cases and work appropriately with them.

The profound effect of outliers on summary measures can be seen clearly in the case of the *mean*. The mean is the arithmetic average of a distribution. Adding together the values of all of the cases in a distribution and then dividing that sum by the total number of cases calculates the mean. One important consequence of this computational procedure is that the value of every case in the distribution enters into the mean's value. And in this sense, the mean can be especially representative of a distribution. Indeed, the mean is sometimes referred to as the "expected value of a variable." That strength, however, is also the mean's weakness. Since the value of every case affects the mean's value, extreme scores have a pernicious effect on its value, either artificially inflating or deflating it, depending on whether they are extremely high or low scores. Thus, in the presence of outliers, the mean does not best describe all of the values in a distribution. The mean is not a resistant summary statistic, and in the presence of outliers it should not be used as a representative or typical value of a distribution. To illustrate this, consider the distributions depicted in Table 3.2.

[Insert Table 3.2 about here.]

In both distributions, the median (the middle score) is 70. There are three modes in these distributions (63, 70, and 71). The mean of Distribution 3.2A is 69.6. The mean of Distribution 3.2B, however, is 221.4. The extreme score recorded for the final case in Distribution 3.2B (9200) grossly inflates the value of its mean. With 59 of the 60 cases having values less than 100, a mean in excess of 200 is clearly not descriptive of the set of scores making up the distribution. If one were to report the mean for Distribution 3.2B

as a typical score, one would draw highly misleading conclusions concerning the scores in the distribution – here, that the losing team typically scored over 200 points.

Clearly, the final case in Distribution 3.2B is an outlier, and given the nature of the concept that the values in this distribution represent (the scores of the losing team in the 1990 NCAA national championship basketball tournament), the outlying case is the result of a processing error. The solution, then, is simply to fix it. That would require examining the source of data again, and replacing the 9200 value with the correct losing score.

Measures of Dispersion

Measures of central tendency report some typical value of a distribution. However, they do not report all the relevant information concerning the values of a variable. Most importantly, by itself, a measure of central tendency offers no indication of how representative its value actually is of a given distribution. (Consider the representativeness of the mean in each distribution displayed in Table 3.2.). Measures of dispersion provide leverage in this regard. They report the extent to which cases differ from one another – that is, how consistent or homogenous the cases are. Consequently, measures of dispersion enable the analyst to assess how representative a given measure of central tendency is. The less dispersed a distribution, the more representative is the value of a measure of central tendency.

It is important to recognize that some measures of dispersion are usable when the values of a variable are ratio (or interval). For nominal or ordinal variable, other measures of dispersion need to be developed

Consider the distributions depicted in Table 3.3. Both distributions have a mean value of 233 and a median of 200. Distribution 3.3A, however, is far less homogenous, and a typical value, whether using the mean or the median, deviates appreciably from several of the scores in the distribution. To give this example some substantive content, let's assume the distributions are the scores of two different political science classes on the same cumulative exam. Both classes had the same average test score – 233, but as a group, the class whose scores are reported in Distribution 3.3B were more consistent in their test performance. To the extent the exam measured a student's understanding of course material, the class whose test scores are reported in Distribution 3.3B attained a

more uniform level of understanding. Consequently, the mean (and the median) is more representative of that class's performance than are these summary statistics for the class whose scores are recorded in Distribution 3.3A.

[Insert Table 3.3 about here.]

Like measures of central tendency, there are several measures of dispersion. The *range*, the *interquartile range*, and the *standard deviation* are among the most widely used. The *range* is simply the difference (distance) between the highest and lowest scores in a distribution. Thus, the range in Distribution 3.3A is 500 – 10 = 490; the range for Distribution in 3.3B is 325 – 180 = 145. The *interquartile range* (IQR) is a similar concept. It measures the difference between the values of the cases at the 75th and 25th percentiles[7] (the upper and lower "hinges," respectively). The IQR for Distribution 3.3A is 125 (290 – 165); the IQR for Distribution 3.3B, 80 (270-190). Both measures (range and IQR) show the spread of a distribution, the bigger the value of the range or IQR, the bigger the distribution's spread. Thus, Distribution 3.3A has a greater spread than Distribution 3.3B, both in terms of the range and the IQR.

This measure of dispersion is sometimes referred to as the "midspread."

In reporting the spread, however, both types of range are rather blunt instruments. First, neither measure takes into account all of the scores of a distribution in its computation. This can be especially problematic in the case of the simple range. It ignores all but two values, the highest and lowest scores in a distribution. Thus, its may be very misleading, particularly in the presence of outliers. Outliers do not affect the IQR, however, since it measures the spread of the middle 50% of a distribution's cases. Here again, however, the computation of the midspread uses only two values, the scores at the 75th and 25th percentiles. Second, neither type of range reports the degree to which any specific value or score in the distribution deviates from some typical or representative score. That level of information is provided by the standard deviation.

[7] The 75th and 25th percentiles in a distribution are the values of the cases below which 75% and 25% of the cases (or observations) in that distribution fall.

The *standard deviation* reports how far, on average, any score in the distribution deviates from the distribution's mean.[8] Thus, the greater a distribution's standard deviation, the more heterogeneous or dissimilar (or spread from the mean) are the cases that compose it. Because the standard deviation is measured in terms of the average deviation from the mean, and the mean is highly vulnerable to extreme scores (outliers), the standard deviation is likewise vulnerable to extreme scores.[9] In short, in the presence of outliers, the standard deviation will give a misleading indication of the average extent to which the values in a distribution are spread around some typical or representative value. To illustrate this, again consider the distributions of NCAA tournament scores presented in Table 3.2. The standard deviation of Distribution 3.2A is about 12. In other words, on average, the scores deviate from the distribution's mean by about 12 points. This measure of average spread is substantially lower than the measure of average spread for Distribution 3.2B. Its standard deviation is about 1200, indicating that, on average, each score deviates from the distribution's mean by almost 1200 points. Yet, the two distributions are identical except for the score of the final case. Relying only upon the standard deviation to formulate a sense of the distribution's spread would result in a grossly misleading interpretation for Distribution 3.2B. The conclusion is that each score in the distribution stands at a substantial distance from every other score in the distribution. In the presence of outliers, then, the value of the standard deviation is not representative of a distribution's spread. Resistant measures are more appropriate in these instances even if the data are interval data, which is required for the calculation of the standard deviation.

---

[8] More specifically, the standard deviation is the square root of the variance. Variance, in turn, measures the average extent to which the scores in a distribution deviate from the mean. Because the mean is the value of a distribution for which the sum of the deviations is 0, variance is computed by summing the squared deviations and then dividing by the number of observations. The standard deviation is a bit more concrete than the variance. By computing the square root of the variance it puts the measure of dispersion back on the same metric as the variable under consideration.

[9] Indeed, since the values of the deviations from the mean are squared in the computation of the standard deviation, the effect of outliers is even more pernicious. This is because scores more distant from the mean not only add to the sum of the squared deviations, they do so at increasing rates. Consider the distribution [2, 4, 6, 8, 10]. Its mean is 6, and its standard deviation is 3.2. Increasing the final value in the distribution from 10 to 100 increases the mean to 24 and the standard deviation to 42.5. The mean increases 4 times; the standard deviation, 13.3 times (see Hartwig and Dearing 1979, 19-20 for a similar example).

As is the case for the measures of central tendency, indices of spread derived from positional measures are more resistant. Hartwig and Dearing (1979) describe a numeric summary of a distribution's spread derived by combining a distribution's extreme scores (its lowest and highest values), the values of the cases at the 25th and 75th percentiles (i.e., the hinges), and its median that is one such measure. To provide a sense of distance, and therefore spread, between these values, a variety of ranges can be computed and reported. One set of ranges reports the distances between the hinges and the extreme values as well as the distances between the hinges and the median. A second set of ranges (referred to as the low-spread, midspread, and high-spread) reports the values between the median and the extreme values as well as the IQR itself. More specifically, the distance between the median and lowest value in the distribution is referred to as the low-spread; the distance between the median and the distribution's greatest value is called the high-spread; the IQR is called the midspread.

Figure 3.1 displays this resistant summary of spread for the NCAA Tournament distributions. If we focus only on the top-half of Figure 3.1 (i.e., the numeric summary for Distribution 3.2A), line A reports the values of the extreme scores, the hinges, and the median. Line B reports the first set of ranges or distances between these values. For example, the distance between the distribution's lowest score and the value of the case at its 25th percentile is 16; the distance between its median and the value of the case at the 75th percentile is 8. Finally, line C
reports the distances of the low-spread, midspread, and high-spread. Again these are distances not data values. They provide some indication of spread that is relatively insensitive to the extremes of a distribution.

[Insert Figure 3.1 about here.]

By reporting the median and the IQR (or midspread), these numeric summaries provide some leverage regarding a distribution's typical value and its degree of spread. The various ranges also allow the analyst to draw some conclusions concerning the distribution's shape. A normal, bell shaped distribution is symmetrical about its representative value, and as Hartwig and Dearing point out, this symmetry is reflected in the near equality of the ranges reported for three pairs of values: (1) the low-spread and the high-spread, (2) the ranges between the hinges and the extreme values, and (3) the

ranges between the median and the hinges (1979, 23). Each of these pairs of ranges in a distribution should be close to identical or equal. Examination of these three pairs of values indicates that Distribution 3.2A is nearly normal. The low-spread and high-spread are 24 and 22 respectively (refer to line C). The distance between the lower hinge and the lowest score and the upper hinge and the greatest score are 16 and 14 respectively. Finally, the distance between the median and either hinge is 8 (refer to line B). This stands in sharp contrast to Distribution 3.2B. Its numeric summary reveals a great deal of inequality among the ranges and therefore asymmetry. The low-spread is 24; the high-spread, 9130. The distance between the lower hinge and the lowest score is 16; the distance between the upper hinge and highest score is 9122. Thus, the outlying case in this distribution makes a great deal of difference even in these measures of dispersion.

Although the numeric summary offers clues concerning the abnormal shape of Distribution 3.2B, it does not provide enough detail to fully understand the cause of it. As Hartwig and Dearing stress, the various summary statistics we have described and discussed are designed to summarize characteristics of the distribution. They do not provide adequate detail to develop a full understanding of the data (1979, 16). A visual representation of the distribution, a graphical image of its shape, provides this level of detail. Truly, a picture is worth a thousand words, or in this case, a picture is worth a thousand numbers!

The Shape of a Distribution

Distributions can have a variety of shapes. They can be bell-shaped and symmetrical (i.e., the classic “normal” distribution). They can be tall and skinny or low and squat, depending on the degree to which the individual cases in a distribution deviate from one another. They can have one hump (mode) or several. And they can be “skewed.” That is, one of the tails (or side) of the distribution is appreciably “longer” than the other. This occurs because the bulk of the observations in the distribution are concentrated at one end (side) of the distribution or the other, leaving relatively fewer cases at the

opposite end, thereby giving the appearance of the distribution's tail being pulled out in the direction of relatively few cases.[10]

Ideally, a distribution would be reported in a way that gives some indication of its shape because understanding a distribution's shape is as important as understanding a distribution's representative values. This is true for at least two reasons. First, the shape of a distribution affects how representative a summary statistic actually is. As we have seen, outliers affect the shape of a distribution, pulling it into asymmetry, and they affect the appropriateness of different measures of central tendency and dispersion. Even in the absence of outliers, however, a distribution's shape affects the representativeness of summary statistics. In skewed distributions, the value of the mean is pulled in the direction of the skew; thus the mean's value is more similar to the values of the relatively few cases at one end of the distribution or the other than it is to the bulk of the cases at the opposite end. The values of the median and the mode, on the other hand, are not similarly affected.[11] Second, knowing the relative frequencies at which the values of a phenomenon of interest occur is important to develop a full understanding of that phenomenon. For example, if we see that the polarization of Congress over time is negatively skewed, we have identified an important pattern in the data, and we might begin to search for temporal correlates (i.e., time-related factors) that might help to account for this development.

Upon the examination of the shape of a distribution and the determination that it is not normal or at least symmetrical, one of "the most powerful of the tools available to the data analyst" is transforming or "re-expressing" the data (McNeill 1977, 12; see also Tukey 1977, chapter 3). The goal here is to mathematically alter the data in the distribution so that they are more symmetrical. Symmetrical

> A distribution is symmetrical if there is a point in the distribution at which every detail on one side is the exact mirror image of every detail on the other side (Chambers et al. 1983, 16).

[10] If the bulk of the cases are concentrated at the high end of the distribution, the distribution is said to be negatively or left skewed. If cases are concentrated at the low end of the distribution, the distribution is positively or right skewed.

[11] Indeed, in a negatively skewed distribution, the value of the mean is less than the value of the median. In a positively skewed distribution, the value of the mean is greater than the value of the median.

distributions have several virtues.[12] First, they are far easier to summarize and interpret because there is no question as to where their center lies. In symmetrical distributions, the center is, simultaneously, the symmetrical center, the 50th percentile, and the mean. In asymmetrical distributions, it is neither clear what the center is nor whether the center even is the most representative value of the distribution. Thus, summary statistics are more representative of symmetrical distributions (see McNeill1977, 12).[13] Second, symmetry simplifies the description of a distribution. One need focus on only one half of the distribution since the other half is identical in every detail. Finally, many traditional statistical techniques and tests are designed to work on symmetrical data.

There are a variety of transformations that can be applied to the data, and many involve changing the original data values by some exponential power. Changing the original values by a power greater than 1 (e.g., computing the square or cube of the value) will reduce the asymmetry of a negative skew. Transforming the original values by a power less than 1 (e.g., computing the square or cube root) will move a positively skewed distribution toward symmetry (Velleman and Hoaglin 1981, 48-49). Log transformations and reciprocals are commonly used as well. In the final analysis there is no set rule for which transformation to apply. Trial and error determines the analyst's choice of how best to re-express the data.

There are a number of ways to display a distribution. The easiest, but also the crudest and least informative way, is to simply present the *raw distribution*. That is, arranging the values of a variable in numerical order and then displaying the ordered value for every observation. The distributions in Tables 3.1, 3.2, and 3.3 are clear examples of just this kind of raw distribution. Raw distributions certainly provide detail with respect to the values of the observations in the distribution, but this type of display does not facilitate an understanding of a distribution's shape. For example, from a raw distribution one cannot tell what the mode of the distribution is or what the median is. Both of these indicators must be determined by counting the values in the distribution. Most tellingly, in a raw distribution, there is no obvious metric to examine that reports

[12] The remainder of this paragraph is based on Chambers et al. 1983, 17-18.

[13] More specifically, this is true for uni-modal distributions. Bi-modal and multi-modal distributions are too complex to be summarized in a single statistic (Hartwig and Dearing 1979, 30).

the relative frequency of the given values. Finally, processing whether or not an outlier is present takes some time. The highest and lowest scores in the distribution must be compared to the next highest and lowest score in the distribution. Then a determination must be made whether or not the distances between those scores is substantial enough to constitute the presence of an outlier.

Frequency distributions partially summarize a raw distribution, and when they are accompanied by percentage and cumulative percentage distributions (as they almost invariably are), they provide a metric to determine the relative incidence of each value in the distribution. Frequency distributions report each value the variable takes on for all the observations in the data set, the number of times each of those values occur in the data, the percentage of times each of those individual values occur in the data, and the running percentage of times a score of a given value of less occurs in the data. Table 3.4 displays the frequency distribution of the basketball scores that appear in Table 3.2, part A.

[Insert Table 3.4 about here.]

From the frequency distribution, we see that there are a total of 60 observations in our data. Many of the losing scores appear in the distribution more than once. For example, four teams lost when scoring 70 points; two teams lost despite scoring 91 points. We can also quickly determine the relative incidence of a given losing score in the data. For example, losing teams that scored 63 points are present in the data 6.67 percent of the time. Finally, we can quickly determine the proportion of cases in the data that have a score of a given value or less. Teams that scored at least 61 points account for 25 percent of the losing teams in the 1990 NCAA tournament.

Clearly, it is possible to wring a fair amount of information concerning the data from a frequency distribution. However, frequency distributions do not lend themselves to an easy determination of a distribution's shape. Careful examination of the cumulative percentages might allow the perceptive analyst to formulate some sense concerning whether a distribution is skewed. Recall that a skewed distribution has a concentration of observations at one end of the distribution or the other. If there is a notable "bulge" in the cumulative percentage at a given end of a frequency distribution, this would suggest skewness. However, this information is not quickly apparent from the frequency distribution. Moreover, the frequency distribution, like the raw distribution, does not

facilitate the identification of outliers. The frequency distribution of the scores from distribution 3.2B would look almost identical to the frequency distribution appearing in Table 3.4. Both distributions have a total of 60 observations. Both frequency distributions indicate that four teams lost when scoring 70 points, and two teams lost despite scoring 91 points. And both frequency distributions would show that teams that scored at least 61 points account for 25 percent of the losing teams. Yet, we know from our preceding discussion, that the shapes of the distributions are very different. One distribution is nearly normal; asymmetry marks the other distribution because of an outlying case (most likely the result of a processing error).

To a significant degree, a *histogram* overcomes a frequency distribution's deficiencies regarding the communication of shape. Indeed, a histogram is a summary graph of a frequency distribution. To construct a histogram, the observations (values) in the distribution are organized into "bins" or categories according to their values. Typically, bins are created so that they contain between five and 20 observations. Once the bins are established, each individual observation is placed in its appropriate bin. The bins are arranged in numerical order along the x-axis. The y-axis reports the number of observations in each bin. A bar drawn for each bin further summarizes this frequency. The bin with the tallest bar has the most observations.

> Histograms for a data set are very sensitive to the "binning" process. There are various decision rules for binning, but it is somewhat depending on one's taste, and will produce much different "pictures" of the distribution, depending on the choice of bins. The usual practice of selecting between five and 20 bins does not provide much guidance for determining whether one should use seven or eight bins for a histogram.

Once constructed, a histogram can be examined to determine a distribution's general shape as well as the presence of outliers. If the bars of the histogram are arranged in a near symmetrical, bell shaped pattern, the distribution is generally normal. If taller bars are concentrated at one end of the x-axis or the other, the distribution is skewed. And if a bin stands well apart from the bulk of the other bins, then this indicates an outlier.

Figure 3.2 depicts a pair of histograms. The histogram on the left portrays the 1990 NCAA losing scores presented in distribution 3.2A. The histogram on the right portrays the distribution of family wealth in the United States. The numeric summary we discussed above (see Figure 3.1) indicated that the losing scores from Distribution 3.2A

are normally distributed, and the histogram of those scores confirms that impression. The histogram is generally bell-shaped and symmetrical. There are relatively few scores at either extreme, and the bulk of the scores fall in the middle.

The histogram of family wealth on the other hand depicts a positively skewed distribution with several outlying cases. The bulk of the distribution is concentrated at the lower end, but there are a few cases with high scores that "pull out" the distribution's tail at the positive end. Moreover, there are two bins that stand well apart from the bulk of the distribution, one between 20 and 30 and the other between 30 and 40. Clearly these are outliers. Furthermore, there are several bins in this distribution that are completely empty.

[Insert Figure 3.2 about here.]

Although histograms communicate the shape of a distribution, they also conceal important information about the distribution, namely the values of the individual cases. Hartwig and Dearing note that a *stem-and-leaf display* combines the numeric information provided by a frequency distribution with the sense of shape communicated by the histogram (1979, 16; see also Tukey 1977, chapter 1). To construct a stem-and-leaf display, the values of the observations in a distribution are ranked and then separated and organized according to their digits. The first digit is the "stem," and it is arranged vertically in ascending order. The subsequent digit is the "leaf." The leaves associated with each stem are arranged horizontally, with leaves of greater values being farther to the right. The stem can be further stretched by subdividing the rows into two rows, one identified with a (•) that includes leaves with values ranging from zero to four and the other identified with a (*) that includes leaves with values ranging from five to nine (Hartwig and Dearing 1979, 17). Figure 3.3 displays the stem-and-leaf plot for the distribution of losing basketball scores. Here again, there is strong evidence that the distribution is close to normal. Moreover, the individual scores are retained. Ninety-two is the greatest score; 46, the lowest. Nearly every value between 46 and 92 is represented in the distribution. In other words, there is no appreciable gap in the losing scores. A histogram, however, conceals this attribute of the distribution because the distance between the individual scores that are placed in each bin is not reported.

[Insert Figure 3.3 about here.]

The stem-leaf display can be completed for any kind of distribution and the stem-leaf of the losing scores (Table 3.2B) show the 9200 score in very striking relief and distance from the remainder of the distribution, far below the rest of the values. Conducting a stem-leaf on data at the very outset will quickly give the researcher a sense (a picture) of the distribution of the variable. It is a very good way to identify extreme values quickly so that they can be corrected of at least checked. Note that the stem-leaf in Figure 3.3 has ten bins. That was automatically determined by the software, rather than set by the researcher. But the result is a very clear picture of the distribution of the lowing scores.

Box-plots can be displayed horizontally or vertically. Figure 3.4 is obviously a horizontal display. In much of what follows the box-plots will be displayed vertically. However, when a horizontal display conveys the information more clearly, horizontal box-plots will be used. The choice of horizontal or vertical is left to the user. Although there may be times when one display better presents a pattern. This means choosing one or the other should be done carefully.

To become very familiar with box-plotting one should consider each box-plot presented in this book in terms of whether a horizontal or vertical display provides the more telling "picture" of the distribution. If there seems to be no difference between the two versions then the choice is left to the taste of the social scientist.

In addition, the outliers, those data values beyond 1.5 times the midspread from the Inner quartiles can be displayed using any symbol. Here, those outliers will be displayed by solid dots. •

Finally, a *box-and-whisker* plot (or box-plot) is, in large measure, a graphical representation of the numeric summary we discussed above. A box-and-whisker plot consists of three horizontal lines composing the box. The lower line represents the value of the distribution at the 25th percentile (the lower hinge). The upper line represents the value of the distribution at the 75th percentile (the upper hinge). The middle line is the distribution's median. The "whiskers" are lines that extend from either end of the box and terminate at the value of cases that are farthest from the hinges. Those are the lowest and the highest data values in the set. The ends of the whiskers are generally no more than 1.5 times the value of the midspread from the lower and upper hinges (the 25th and the 75th percentile values). However, that is a matter of preference on the part of the researcher, and some prefer the whisker to extend no more than one midspread from either hinge. Any case that has a value that results in it departing by more than one or 1.5 times the IQR from either hinge is marked individually with a (•). These values are considered to be extreme values and warrant some individual attention

To assess the symmetry of a distribution by examining a box-plot of the data, both the shape of the box and the length of the whiskers need to be examined. That is because the shape of a distribution depends on both the shape of the box (and the location of the median in the box) and the length of both whiskers.

on the part of the researcher both in terms of explanation and statistical treatment. Figure 3.4 displays a box-and-whisker plot (box-plot) of the 1990 NCAA tournament losing scores from Distribution 3.2A.

[Insert Figure 3.4 about here.]

A box-plot quickly communicates a graphic sense of a distribution's shape. The location of the median line relative to the upper and lower hinges indicates the degree to which the inner quartiles are normal. If the median is fairly centered between the hinges, then the distribution is normal. If the median is appreciably closer to the lower hinge, the distribution has a positive skew, and if the median is closer to the upper hinge, the distribution is negatively skewed. The length of the whiskers is also very important for getting a sense of the distribution of the variable since the whiskers represent data values in the two outer quartiles. Whiskers of about equal length confirm the sense of the normality of the distribution. If one of the whiskers is very short then that tail of the distribution suggests a skewed distribution even if the box is fairly symmetrical. The presence of symbols for some of the data points is a quick and clear indication of whether outliers are present. That means it is important to have a clear sense of the "distance" from the hinges (lower and upper quartiles) where the whiskers end and symbols start appearing. The standard 1.5 IQR measure is used most often for this purpose. The box-plot in Figure 3.4 indicates that the losing scores are very nearly normal. Moreover, there is no symbol appearing beyond either whisker.

We have determined in earlier discussions of this distribution that its shape is normal and there are no outlying cases. One matter that has not been emphasized to this point is the vertical scale in Figure 3.4. In comparison with the whiskers, there is the clear evidence that the lengths of both whiskers are nearly identical. If the scale in examined and one moves 20 points from the median (70), one can tell that the losing basketball teams might have scored a few less points on the lower end than they did at the upper end. However, there is very little difference in the length of the two whiskers more than 20 pints out (either direction) from the median.

The box-plot is an efficient and quick way of examining a set of observations (a variable) and determining the shape and central location of the distribution in one graphical step. These plots are quite useful and they will be used in subsequent chapters to analyze data and illustrate their great value in uncovering the mysteries of social science data. These techniques can be used on all kinds of data, although they are most useful for interval data, as are the more sophisticated measures of dispersion outlined earlier in this chapter.

The median of the distribution is clearly displayed in the box-plot. The median can also be displayed in a stem-leaf plot by either highlighting the median value or marking its locating with a vertical line. The box-plot does not display the location of the mode of a distribution because it is a graphical display of distances between values in the distribution. One could "see" the mode(s) of the losing scores in the stem-leaf by observing that there are four threes after the 6* in the stem, and four zeroes and one's after the 7* in the stem.

Conclusions

This discussion has focused on both traditional and exploratory (descriptive) methods of assessing the centrality, the shape, and the spread of a single variable. It is very important to undertake this kind of analysis for each variable in a data set. If a variable does NOT vary (as the year of the case filings outlined in the previous chapter) then there is obviously no need to undertake any of this kind of analysis. (The year will become important if, later case filing data are collected for additional years, beyond 2005. That is because it then becomes a variable rather than merely an indentifying characteristic of all the district court case filings.)

The value of these techniques is to permit quick and efficient assessment of the characteristics of each variable in a data set. Later analysis will focus on comparisons among variables and efforts to link variables or assess the relationships that may exist between or among several variables. However, before that can be done, it is important to answer the questions about the "mystery" of each variable, so that the researcher has a very well-developed understanding of the data they are working with.

What follows in the next chapter will involve the use of these descriptive methods – exploratory methods – to make comparisons among a real set data. Stem-leaf plots and box-plots can be quite useful for comparing variables and categories of variables. The

existence of patterns in the variables, the likelihood of connections or relationships, and the nature of those are important for the social scientist to uncover and these techniques permit a great deal of the mystery of data to be uncovered.

Table 3.1. Two Hypothetical Distributions.

Distribution 3.1A	Distribution 3.1B
10	10
18	15
22	22
23	23
24	40
50	41
65	

Table 3.2. Losing Scores, 1990 NCAA Basketball Tournament.

Table 3.2A				Table 3.2B		
46	67	83		46	67	83
47	68	83		47	68	83
48	70	84		48	70	84
52	70	85		52	70	85
52	70	86		52	70	86
53	70	88		53	70	88
54	71	89		54	71	89
54	71	91		54	71	91
55	71	91		55	71	91
56	71	92		56	71	9200
58	72			58	72	
60	72			60	72	
60	72			60	72	
61	73			61	73	
61	73			61	73	
63	75			63	75	
63	75			63	75	
63	77			63	77	
63	78			63	78	
64	78			64	78	
65	78			65	78	
65	79			65	79	
66	80			66	80	
67	81			67	81	
67	81			67	81	

Table 3.3. Sample Test Scores from Two Classes.

Class A	Class B
10	180
165	190
200	200
290	270
500	325

Table 3.4. Frequency Distribution of Losing Scores, 1990 NCAA Tournament.

Losing Score	Freq.	Percent	Cum.
46	1	1.67	1.67
47	1	1.67	3.33
48	1	1.67	5.00
52	2	3.33	8.33
53	1	1.67	10.00
54	2	3.33	13.33
55	1	1.67	15.00
56	1	1.67	16.67
58	1	1.67	18.33
60	2	3.33	21.67
61	2	3.33	25.00
63	4	6.67	31.67
64	1	1.67	33.33
65	2	3.33	36.67
66	1	1.67	38.33
67	3	5.00	43.33
68	1	1.67	45.00
70	4	6.67	51.67
71	4	6.67	58.33
72	3	5.00	63.33
73	2	3.33	66.67
75	2	3.33	70.00
77	1	1.67	71.67
78	3	5.00	76.67
79	1	1.67	78.33
80	1	1.67	80.00
81	2	3.33	83.33
83	2	3.33	86.67
84	1	1.67	88.33
85	1	1.67	90.00
86	1	1.67	91.67
88	1	1.67	93.33
89	1	1.67	95.00
91	2	3.33	98.33
92	1	1.67	100.00
Total	60	100.00	

Figure 3.1. Examples of Distribution Indicators.

Distribution 3.2A

46 62 70 78 92 [A]

16 8 8 14 [B]

24 16 22 [C]

Distribution 3.2B

46 62 70 78 9200

16 8 8 9122

24 16 9130

Figure 3.2 Two Examples of Histograms.

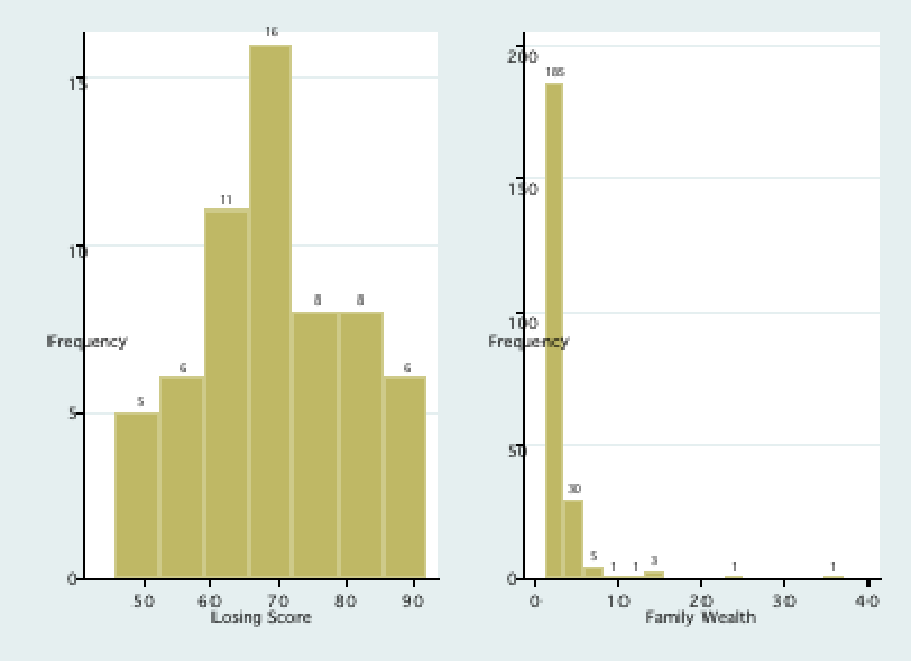
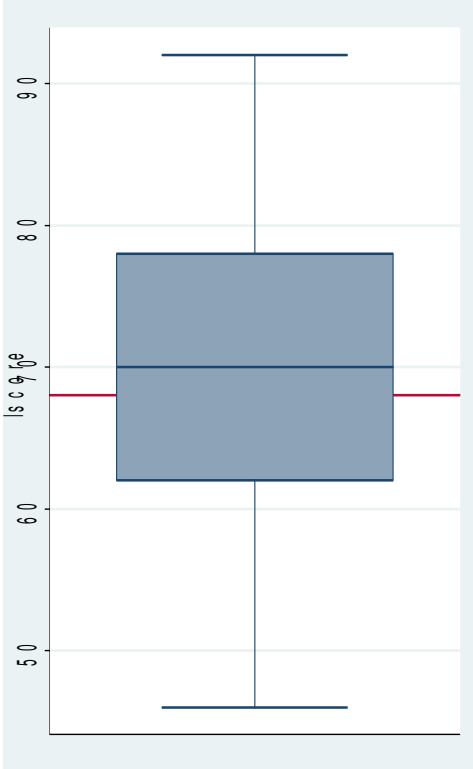


Figure 3.3. Stem-and-Leaf Display of Losing Scores.

```
4.  | 678
5*  | 22344
5.  | 568
6*  | 001133334
6.  | 5567778
7*  | 0000111122233
7.  | 5578889
8*  | 011334
8.  | 5689
9*  | 112
```

Figure 3.4. Box-and-Whisker Plot (Boxplot), 1990 NCAA Tournament Losing Scores.



CHAPTER 4

A Univariate Application

Since the general features of shape, spread, and location of a data set were all discussed in the abstract in Chapter 3, it is important to present some real data and to examine the spread, the shape, and the central locations of those data. Using the data source discussed in Chapter 3 (federal district court case filings) and the techniques that were introduced there (stem-leafs and box-plots), the focus here to examine the location, shape and spread of the 89 federal district courts in the United States. That workload was measured by examining the total number of cases filed in each district, and to examine one year of these data – 2005.

Exploratory Data Analysis of District Court Case Filings

The court filings ranged in size from 486 to 17099 cases during 2005. The mean for these was 3438, while the median was 2198. Since the mean and the median are nowhere nearly identical, the location of these data is not easily and accurately determined by one number. Furthermore, this disparity indicates that the distribution is asymmetrical. The median and the mean of a perfectly normal distribution are identical. The difference between the mean and the median (1240 cases) is quite large in comparison with the value of these two indicators on centrality, the amount of skew in the distribution might be quite great.

Examining the stem-leaf display for these data is the next step and that indicates a great deal of positive skewness in the case filings data. Figure 4.1 displays the stem-leaf for the raw data. The shape and the spread of the case filings is clearly not normal and or even symmetrical. A visual assessment of the stem-leaf is essential in order to "read" the distribution it displays. Reading the stem-leaf requires some attention. It is necessary to know how the stem-leaf was constructed. The stem in Figure 4.1 (the number to the left of the vertical line) refers to the thousandth digit. This needs to be determined by actually inspecting the data and the stem carefully. The smallest number in the stem 0*** refers to a zero in the thousandth digit. The smallest number of case filings is 486. There is one district court had 486 case filings in 2005 and that was the smallest filings of any district.

Moving down the stem, one can determine that the first 2*** value in the stem involves case filings that numbered at least 2000. The leaves for this segment of the stem (after the first "2" and to the right of the vertical line) finish off the numbers. These should be read as: 2001, 2110, 2190 (the median), 2223, 2297, and 2424. The rest of the case filings in the 2000s (the second line is also marked 2***) are all above 2500. Their values are 2697, 2785, and 2990. The full range of cases from a minimum of 486 to a maximum of 17099 is quite staggering. The first item of note here is the large maximum, 17099 and the other notable extreme values involving case filings that were greater than 10,000 cases.

[Insert Figure 4.1 about here.]

The immediate question is whether these extreme high values are correct or mistakes of data entry. So, the next step is to check the original data source to see whether the values were correctly entered in the data file. To do that, the districts that are attached to these filings need to be identified. It turns out that these values were derived from the Eastern District Court of Pennsylvania (17099), the Central District of California (13834), the Southern District of Texas (13332), and the Southern District of New York (12545). When these numbers are checked in the source of the data, it turns out that they are correct values. So there are four very noticeable districts among the 89 districts with very high values. These values clearly illustrate how a few values in a distribution can pull up the mean. These high values do not have the same affect on the median. That is because the median is indicates location, and location is dependent only on determining the middle value. That does not require any calculation of a value.

These extreme values will require some special attention. We could drop these values out of the analysis because they are so much larger than the rest of the data set. (The next highest value is 8859, which is much lower than the 12545 value for the Southern District of New York.) If we drop these values from the analysis then each of them warrants unique and separate attention, analysis, and explanation. However, that would change the shape and the distribution of the data somewhat. The first step here, however, might be to bring them into the analysis, by means of transforming the data. (Transformations of these data will be discussed below.) Transformations of the data change the values of the data but they do not change the location of each item in the data. That means that the locational indicators – median, hinges, and extremes – do not move

or change locations. The highest value will remain the highest value in the distribution. It is very important to remember that about transformations.

It is important to remember that transformations of data can accomplish one of two objectives. As used in this discussion, the transformation will be used to move the asymmetric distribution of case filings toward a more symmetrical (not necessarily normal) distribution or shape. Changing the shape of the data by transforming the data does not move the relative location of any of the data. It does change the value of the mean and the median, as well as the value of the extremes, the hinges and the midspreads. However, those changes affect the shape and the spread of the distribution. The location of the median and the hinges do not change when data are transformed. That is what will be sought in this discussion of district court case filings. (The other objective of a transformation is to straighten out a curve. That use of transformations will be discussed and demonstrated later.)

Before we seek to transform the data, we explore the possibility of dropping the extremely high values in the data set. The distribution of the data, when these four are dropped out, is shown by the stem-leaf in Figure 4.2. The distribution is still skewed. While the highest four values are gone, now there are six districts that still stand out, with case filings in the 8000 range. These are visible in the first stem-lead (Figure 4.1). So dropping the highest outliers only changes the problem of high or low values. It is possible that dropping a set of high values in a data set will make the rest of the data more compact. That does not happen here, since the next six values are still separated from the rest of the data. Dropping values from a set does not change the shape of the distribution. So removing data does not achieve the objective of making a data set more symmetrical. So on first impression, it seems as if the best treatment of the outliers among the 89 districts should be some kind of transformation of the data.

[Insert Figure 4.2 about here.]

Examining a box-plot of the full data set shows graphically just how skewed and what the shape of such a box-plot looks like. One should always examine a box-plot of the data. It shows the location of the hinges and the median. And box-plots indicate the presence of any outliers in the data. The box-plot of the entire data set is displayed in Figure 4.3. A normal distribution produces a symmetrical and a proportional box and

equally long whiskers above and below the box, with no outliers. Clearly that is not the case with these data.

[Insert Figure 4.3 about here.]

Visually, it is obvious that some of the high case filings are outliers. However, it is also clear from the box-plot that the number of outliers is more than must the highest four filings that the stem-leaf displayed in Figure 4.1. The box-plot clearly shows how far out the outliers are and what the rest of the case filing distribution looks like. There are several important features of the box-plot that need to be emphasized besides its shape. The location of the median, inside the box shows just how "low" the median is in comparison with the outliers and the upper half of the distribution. The mean is also plotted, as a separate, independent (red) line on the box-plot and its location is exceptional, largely because the high outliers pull the mean up. Although the mean is inside the box, it is nearly 1300 cases larger than the median, and that distance is quite striking when displayed in the box-plot.

In a symmetrical distribution several features of the box-plot would be different. The shape of the box, the location of the median, and the length of the whiskers would be even on both sides of the central locator (the median) in a symmetrical distribution. That yields several indicators of symmetry that will be used here. The formulas below would indicate a symmetrical distribution. The degree to which the formulas produce numbers other than those indicated by the formulas, the distribution is not symmetrical. The formula generates a ratio (called the Inner Ratio) between the two inner quartile values or the midspread (upper hinge – lower hinge) and the second quartile value (median – lower hinge).[14] These ratios can range from zero to 1.0. The IR of a symmetrical distribution will be at or approach 0.5. How far a distribution is from symmetrical can be assessed by calculating these ratios. And a transformation that moves the ratios toward 0.5 is an improvement.

(median – lower hinge) / (upper hinge – lower hinge) = .5 (Inner Ratio or IR)

This calculation indicates how closely the inner quartiles are to symmetry.

[14] This indicator is taken from McNeil (1977, 38) and the discussion about the development of ratios presented there.

As an example, the following hypothetical indicators show a symmetric distribution:

Low value = 13

2 Q (lower hinge) = 22

Median = 37

3 Q (upper hinge) = 51

High value = 60

The formula above produces the following calculation:

(37 – 22) = 15

(51 – 22) = 29 (midspread)

(15/29) = 0.517

The fact that this ratio is > 0.5 means these hypothetical data are skewed slightly left. A ratio that is lower than 0.5 is skewed right. This hypothetical illustrates a distribution that is very close to symmetrical within the middle two quartiles. The raw data on case filings indicate just now asymmetrical this distribution is. The Inner Ratio (IR) for the case filings is 0.276. This ratio is somewhat rounded but it indicates a substantial amount of right skewing in the raw case filings data.

A second ratio, here called the Outer ratio or OR can be calculated measuring the relation between the owe half of the data distribution and the full range of the data. The symmetrical dat set would have an OR of 0.5 just like the IR. There is no necessary correspondence between the IR and the OR. However, comparisons of these two ratios are helpful in assessing the skewness of the entire distribution. In the case of the Case Filings data, the OR is calculated below:

(median – lowest value) / (highest value – lowest value) = OR

For the case filings this calculation produces an OR as follows:

(2198 – 486) / (17099 – 486) = 1712/16613 = 0.103

This ratio which is further from the target of 0.5 than the IR indicates that the extremes or the tails of the distribution have am even greater affect on the asymmetry than do the inner portion of the distribution (the IR which is inside the box in the box-plot).

The outliers are marked as solid circles above the upper whisker in Figure 4.3. These are calculated as being more than 1.5 times the midspread (the distance inside the

box). The determination of which case filings are **outliers** in the box-plot involves: (1) calculating the midspread, (2) multiplying it by 1.5, (3) adding that value to the value of the upper hinge, and (4) determining which case filings exceed that value. The following reflects the calculation of these for the raw case filing data:

Upper Hinge = 4259

Lower Hinge = 1416

Midspread = 2843 (* 1.5) = 4264.5

4259 + 4264.5 = 8523.5

A case filing value greater than 8523.5 qualifies a district court as an outlier in this data set. That actually yields a total of eight districts with outlier values. There are no separate dots on Figure 4.3 for each of these districts because some of these extreme filings are nearly identical and so their placement in the figure largely overlaps one another. Figure 4.1 also indicates that some (four) of these outliers are very near to the top of the upper whisker, and just barely qualify as outliers using the (1.5 * midspread) calculation. Table 4.1 lists the outlier districts and their values. This is largely for substantive purposes because eventually these outliers will need to be explained. Examining the table and the particular courts might provide some suggestion for why these districts had such large case filings. However, here the focus is on identifying their existence rather than explaining the filings in these districts.

To determine low extremes, the 1.5 multiple of the midspread is subtracted from the lower hinge. There are no lower extremes in these data.

[Insert Table 4.1 about here.]

Dropping the highest eight districts' filings (the outliers) does little to change the shape of the distribution. It does change the dispersion of the data set. That really only tells us that the outliers have been removed. If a stem-leaf plot is done of the 81 remaining districts, the shape of the distribution would be largely identical to the display in Figure 4.1. (Figure 4.2 displays the distribution without the highest four districts. It illustrates that dropping those highest outliers – the highest four – does nothing for the distribution or it shape, except to remove the very high values in the distribution. Dropping all eight outliers does little more for the distribution.) Dropping high (or low) values in a data set changes the dispersion of the data and the shape of the data set.

However, without the four or eight high values that are outliers, the information about those districts is lost to the analysis, unless one analyzes those separately.

Since the shape of the distribution is still skewed, the alternative to dropping values for this variable (case filings) is to try transformations in order to change the shape of the distribution (and thereby to generate IR and OR ratios that approach 0.5). The three characteristics of a set of data are the central or typical value, the dispersion of the variable, and the shape of the distribution. The first item to note is that ALL the values are included when you transform these data. The second point to remember about this transformation is that the objective is to make the data more symmetrical, if not normal. One of the functions that transforming data should do is "to reel in" or bring down (or up) the extreme values and make the distribution accommodate those.

Transforming Data

Transformation was suggested in Chapter 3. The purpose of transforming a single variable is to try to make the variable symmetric or at least more symmetric than the raw data are. This purpose is not just to make the data "look" nicer. Rather, the purpose of transforming the data is also to minimize (if not eliminate) the difference between the median and the mean for the data set. Such minimization of the difference will yield a more symmetrical distribution than the original data. Rough indicators of symmetry are the IR and OR ratios that were explained earlier in this chapter, so a transformation that moves the ratios closer to 0.5 can generally be considered an improvement over the raw data. There is no clear rule about how precise (close) the ratio has to be to the ideal 0.5, so this is ratio indicator is a matter of taste as much as a statistical indicator.

The term that will be used throughout this discussion is to "transform" the data. Tukey (1977, Ch. 3) discusses "re-expression." That may be the preferable term since the values in a data set are being re-expressed by some procedure. The objective, in both cases, whether re-expressing or transforming data, is to make the distribution more symmetrical. See also McNeil (1977, in passim) and Mosteller and Tukey (1977, Chs. 4 and 5, and Appendix.

There are several different kinds of transformations. One transformation would be to add a constant to all the data values. This does nothing for the shape of the distribution, so it will not be considered here. The other kinds of transformations are logs or powers such as squaring all the data values or raising the data to a fractional power. In addition,

to these multiplicative transforms, another set of re-expressions involves taking the reciprocal of the raw data. This is a "divisional" rather than a multiple transformation. There is an order to the transformations that Mosteller and Tuckey (1977, 79-81) call a "ladder." The ladder, from slightest transformation to most powerful, depends on the shape of the raw data and how much "correction" is necessary to improve the shape and the distribution of the data. For our purp0oses, given the shape of the court filings, the following "ladder" contains the possible transformations that could improve the shape of the distribution:[15]

Raw Data

Square Roots

Cube Roots

Fourth Roots

Logarithms

Reciprocal Square Roots

Reciprocals.

Raising a number of the ½ power involves taking the square root of the number. The $^{1/3}$ power is the cube root.

As already noted, the choice of a transforming procedure involves some guess-work, and it does depend on how skewed the Raw Data is. Given the stem –leaf (Figure 4.1) and the box-plot (Figure 4.3) the skew of the court filings is pretty severe. So it may not be worthwhile choosing one of the "milder" transformations such as the square root or the cube root.

Tukey (1977) recommends the use of either a square root or a log transformation.[16] These are the most likely ones to move the data toward a symmetric distribution and move extreme values closer to the bulk of the data. In addition, one of the objectives of transforming the data is to "move" the median and the mean closer together. Re-expressing the case filings as their log produces the stem-leaf displayed in Figure 4.4.

[Insert Figure 4.4 about here.]

[15] This particular ladder is taken from McNeil (1977) at 38.

[16] Tukey(1977, Ch. 3) indicates that the base of the log is not particularly important since any log base only changes the proportion of distance between values. For convenience here, log base 10 ($\log_{(10)}$)will be used.

The first result from looking at Figure 4.4 is that the four, visibly high values are "still there." However, in relation to their location in the raw data the transformation has brought them closer to the rest of the data. However, they are still disconnected from the rest of the transformed data. So even if we stop with this transformation, some individual attention needs to be devoted to explaining or accounting for the case filings in those four districts with the highest values that are listed in Table 1.1. To read this stem-leaf the actual number 26*|9 is the log(10) of 2.69. That means that the minimum number of case filings (raw number = 486) is $10^{2.69}$.[17] That may have no intuitive meaning for a social scientist, but it provides a much different (improved) picture of the case filing data.

The second feature of the transformed data that is apparent in Figure 4.4 involves the spread of the data. The shape of the transformed data has changed. Except for the four high values that remain the spread is much more compact and symmetrical than the raw number of case filings. The shape of the distribution has changed a good deal from the original data. That does not mean the Log of the case filings is a normal distribution. However, it is somewhat closer to normal. The symmetry of the log is another, important question, and the box-plot, displayed in Figure 4.5, illustrates that symmetry better than the stem-leaf.

[Insert Figure 4.5 about here.]

First, despite the existence of those four high districts, they are no longer outliers in the EDA sense! There are no low or high extremes for these transformed data. That is a distinct improvement over the original data, which contained eight, high outliers. (See Figure 4.3.) The second point to note about this box-plot is that the whiskers are nearly the same length. The upper whisker is a bit longer than the lower one, and that reflects the difference that is displayed in the stem-leaf in Figure 4.4. However, the difference in the whiskers is not very great. Another point of interest in the box-plot is that the median is located nearer to the center of the box (the Inner Quartiles) than in the box-plot of the raw data. That also means that the transformation of the raw data has improved this indicator of the typical or central point in the distribution. Lastly, the mean of the tranformed data, the red line on the edges of the box, is closer to the center and the

[17] This kind of exponent might best be understood by remembering that $10^2 = 100$, $10^3 = 1000$, and $10^4 = 10000$. That means that the values of the $\log_{(10)}$ for these raw data extend to nearly $10^{4.3}$.

median of the box than in the original, raw data box. The distance between the mean and the median has narrowed as a result of the transformation.

As was noted above, selecting a transformation function is a matter of trial-and-error or a matter of taste and a more powerful transformation might move the distribution even closer to the ideal IR and OR ratios. So, if we move down the transformation ladder to the bottom, we will see if the Reciprocal of the raw data improves the ratios even more than the log.

A transformation of the data does not allow the comparison of the values for the raw data and the transformed data. However, the ratios for the raw data and the transformed data (for both the Log and the Reciprocal) that were explained above are contained in Table 4.2. Comparing these ratios provides an indication of how much the transformation improved the symmetry of the data. It also indicates just how much the transformation process is a matter of choice and taste. The Inner Ratio is improved for the Log transformation and the Reciprocal transformation also improves the IR. The IR for the reciprocal is very close to 0.5 and would seem to be the transformation of choice. However, the OR for the Reciprocal is hardly better than the raw data (0.199) versus 0.103), while the Log OR is very close to 0.5.Tthat means the choice of transformation is up to the analyst. It is possible that stem-leafs and the box-plots of these two d=transformations can be compared and some choice of transformation can be made on the basis of appearance.

[Insert Table 4.2 about here.]

The stem-leaf and box-plot of the Reciprocal display the result of this transformation and the Reciprocal seems less satisfactory than the $Log_{(10)}$. These are displayed in Figures 4.6 and 4.7. The results are striking because they did not produce a symmetric distribution at all. There are a number of high outliers in this display of the Reciprocal. That confirims the OR for the Reciprocal which is so much lower than the Log OR. The reason the Inner Ratio is improved by the Reciprocal is that only the inner two quartiles and the median are used in calculating that ratio, and so the first and fourth quartiles and the values of that are NOT considered.

[Insert Figure 4.6 and 4.7 about here.]

Figure 4.7 displays two box-plots, the one on the left is the box-plots for the Reciprocal transform. The one of the right is the Log transformation of the same data. Clearly the comparisons of these two indicate how valuable the visual inspection of the data can be. This comparison (called Comparative Box-Plots) also allows you to assess which of the transformations is the most useful for our purposes of making the data more symmetrical. There is no doubt which transformation does that.

The conclusion to be drawn from this brief discussion of transformations is that the picture or the graphic display is at least as valuable as a calculated indicator is for determining symmetry of a distribution. It is essential (and easier), when engaging in Exploratory Data Analysis to graph the data using an appropriate method of display. Examining the graphs, one can obtain a "picture" of the data and what they represent. The result of looking at the figures (stem-leaf and box-plot) for each of the transformations leads to the conclusion that the $Log_{(10)}$ transform is much more satisfactory in generating a symmetrical distribution for the district court case filings, than is the Reciprocal. The two stem-leafs (Figures 4.4 and 4.6) are equally as useful as the box-plots for determining which of the transformations produces the more symmetrical distribution. Perhaps this simple comparison confirms that a picture is worth a thousand words or numbers.[18]

Conclusion

This discussion or example of exploratory data analysis of a set of real-world data may seem simplistic. However, it is important to realize that EDA is relatively simple and it is visual. It is also a very revealing approach to understanding a set of data. **Looking at** the stem-lead and the box-plot for a variable can tell you a good deal about the nature of the variable. It may also provide you with highlights, such as outliers that should be explored individually (after they are checked for accuracy). Visual inspection of a display may seem simplistic and not rigorous. However, there is no substitute for such an examination. In addition, examining such graphics should provide the analyst with a very good sense of what is transpiring or what the shape of the data is.

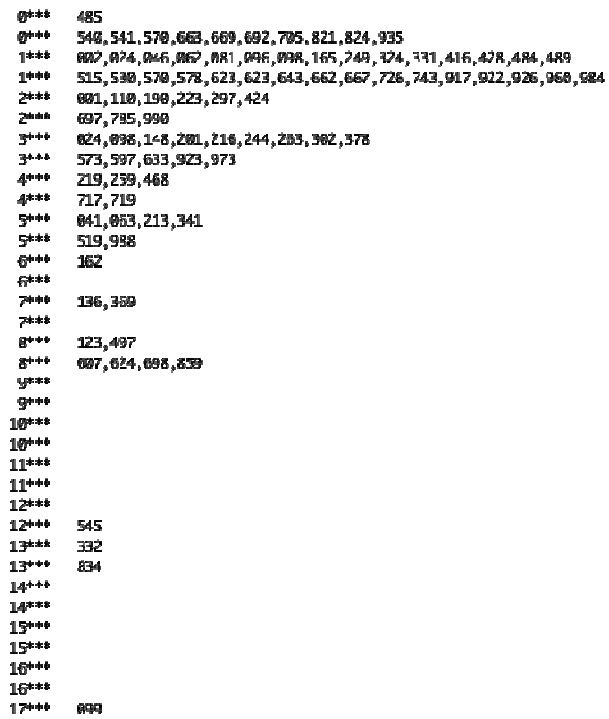
---

[18] In fact we only have 89 number of this data set, but the box-plots are obviously of very great value.

The $Log_{(10)}$ transformation performed on the data seems produced a satisfactory result. Certainly the result is not perfect. However, it is a substantial improvement over the raw data for a variety of reasons. That improvement is confirmed by the plots of the $Log_{(10)}$ and the examination of the IR and OR. Furthermore, the alternative transformation – the Reciprocal – proved to be much less satisfactory than the Log of the Case fFlings. As a result of this exercise, several portions of the analysis that follows will rely on the transformed data rather than the raw data.

The routines that were used to examine the district court filings were software commands that produce displays of the data values or the transformed data. That is necessary in order to make determinations about the values of the range, whiskers, and IQR. Along with determining the value of the median and mean, these values are essential to assessing whether the raw data or some transformation of the data produces the kind of typical value and dispersion and shape to a variable that improves our understanding. We have not examined the question WHY these patterns existed in the data. The detective work is hardly done at this point. That will come later in this material. However, it is important for the social science detective to know first what the facts are – what happened. Then we can begin to explore WHY it happened.

Figure 4.1 Stem-Leaf Plot of Federal District Court Filings, ,2005.



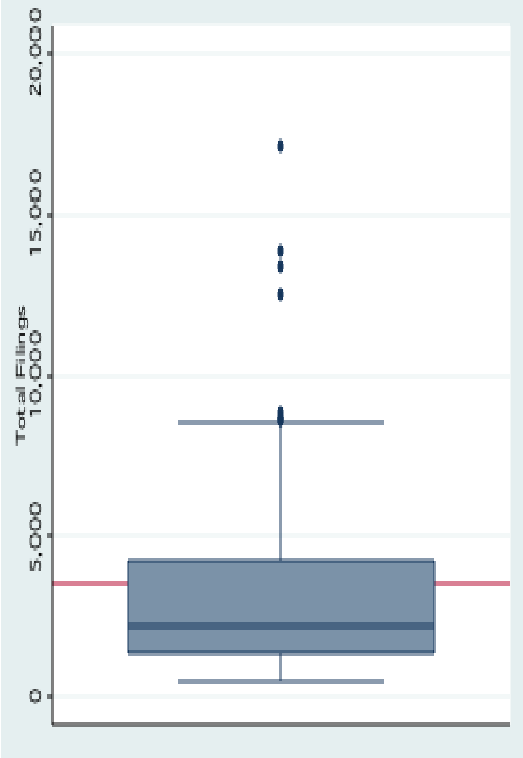
Source: Federal Court Management Statistics, Admin. Office of the U.S. Courts (2005).

Figure 4.2. Stem-Leaf display of 85 Federal District Court Case Filings, 2005.

0*** | 485
0*** | 540,541,570,663,669,692,705,821,824,935
1*** | 002,024,046,062,081,096,098,165,249,324,331,416,428,484,489
1*** | 515,530,570,578,623,623,643,662,667,726,743,917,922,926,960,984
2*** | 001,110,190,223,297,424
2*** | 697,785,990
3*** | 024,098,148,201,216,244,263,302,378
3*** | 573,597,633,923,973
4*** | 219,259,468
4*** | 717,719
5*** | 041,063,213,341
5*** | 519,988
6*** | 162
6*** |
7*** | 136,369
7*** |
8*** | 123,497
8*** | 607,624,698,859

Source: Federal Court Management Statistics, Admin. Office of the U.S. Courts.

Figure 4.3. Box-plot of Federal District Court Filings, 2005.



Source: Federal Court Management Statistics, Admin. Office of the U.S. Courts.

Table 4.1. District Court Outliers and Case Filings, 2005.

District	Case Filings
Pennsylvania Eastern	17099
California Central	13834
Texas Southern	13332
New York Southern	12545
Illinois Northern	8859
Florida Southern	8698
Texas Western	8624
Ohio Northern	8607

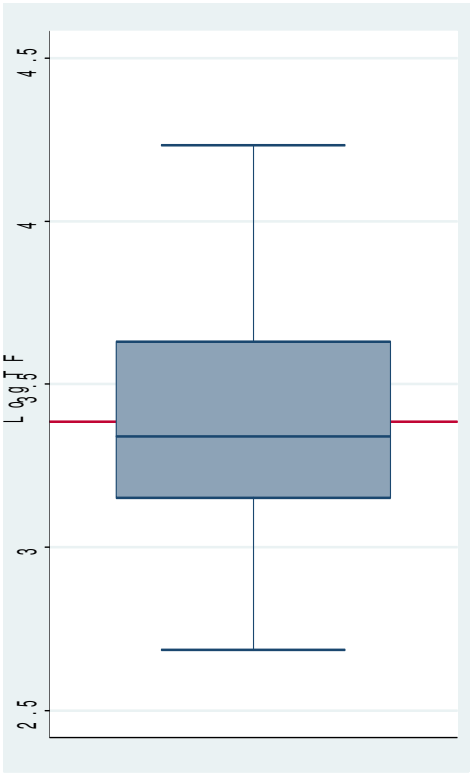
Source: Federal Court Management Statistics, Admin. Office of the U.S. Courts.

Figure 4.4. Stem-Leaf of the $Log_{(10)}$ of District Court Case Filings , 2005.

```
26* | 9
27* | 336
28* | 2345
29* | 127
30* | 01233447
31* | 022557788
32* | 0011222448889
33* | 0024568
34* | 34889
35* | 01111235669
36* | 033577
37* | 0023489
38* | 57
39* | 133445
40* |
41* | 024
42* | 3
```

Source: Federal Court Management Statistics, Admin. Office of the U.S. Courts.

Figure 4.5. Box-plot of $Log_{(10)}$ District Court Case Filings, 2005.



Source: Federal Court Management Statistics, Admin. Office of the U.S. Courts.

Table 4.2. Comparison of Dispersion Ratios for Various Transformations.

	Raw Data	Log of data	Reciprocal of Data
Inner Ratio (IR)	0.272	0.399	0.471
Outer Ratio (OR)	0.103	0.423	0.199

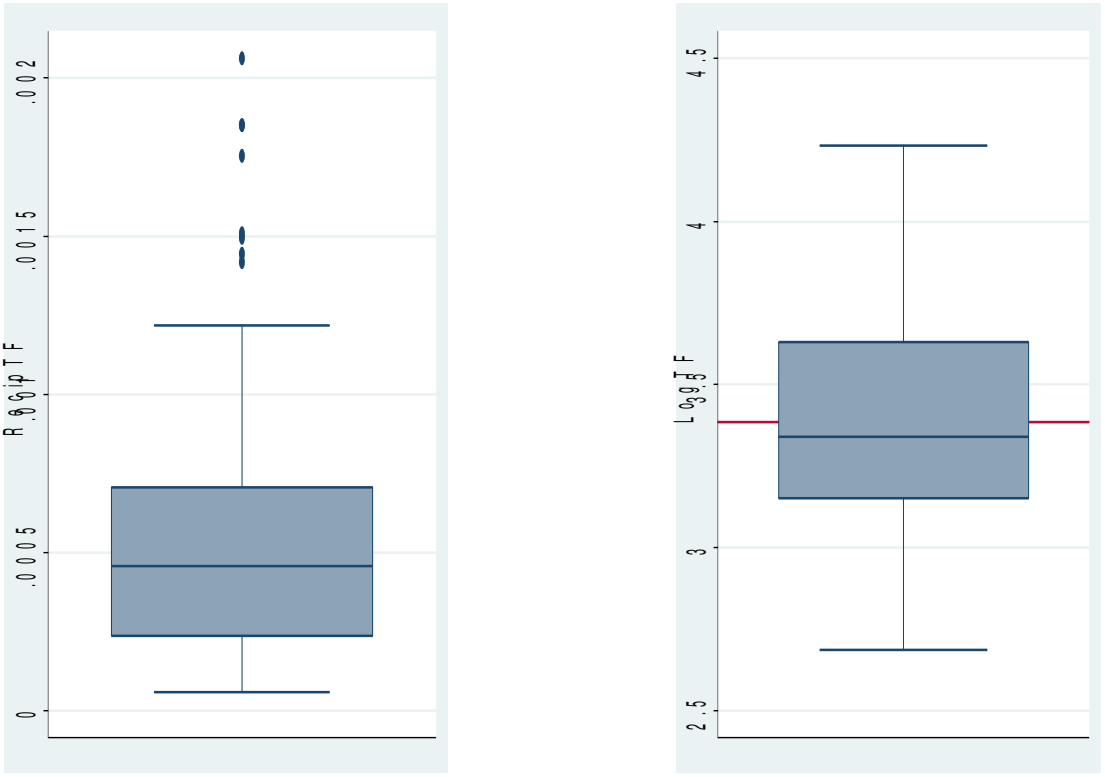
Figure 4.6. Stem-Leaf of Reciprocal of District Court Case Filings, 2005.

```
 0* | 6788
 1* | 1122224467899
 2* | 001123455888
 3* | 001111223367
 4* | 14567
 5* | 00122278
 6* | 00122345677
 7* | 0156
 8* | 06
 9* | 113468
10* | 07
11* |
12* | 12
13* |
14* | 259
15* | 1
16* |
17* | 5
18* | 55
19* |
20* | 6
```

The stem of this plot represents a multiple of 0.00006 or0.00206.

Source: Federal Court Management Statistics, Admin. Office of the U.S. Courts.

Figure 4.7. Comparative Box-plots of Reciprocal and the $Log_{(10)}$ Transformations of the District Court Case Filings, 2005.



Reciprocal $Log_{(10)}$

Source: Federal Court Management Statistics, Admin. Office of the U.S. Courts.

Appendix for Chapter 4.

The analysis completed in this Chapter was done using STATA software. The commands were performed on one variable called `totalfilings`.

The STATA commands follow.

Figure 4.1

```
stem totalfilings
```

Figure 4.2 [This figure was produced by the same command as Figure 4.1 – stem – except that the four high values were dropped by copying the variable `totalfilings` and then replacing those values with the missing value indicator.]

Figure 4.3

```
gr box totalfilings, yline(3438.7)
```

[This command generated the box-plot and the line (yline) that displays the mean for the case filing data as a separate horizontal line. The mean and other statistics for the raw data were produced by the "summary" command.

```
su totalfilings, d
```

[The output of this command is a set of numbers that indicate the mean, the median, the inner quartile values, the extreme values and from this set of numbers the range, and other values in the raw data can be calculated. The "d" in the summary command is essential to produce the necessary values for the variable. It stands for a "detailed" summary.]

To create the log of the variable the following generate command was entered. The gen command creates a new variable, here called `LogTF`.

```
gen LogTF = (log10(totalfilings))
```

Figure 4.4

```
stem LogTF
```

Figure 4.5

```
gr box LogTF, yline(3.384)
```

[As with Figure 4.3, this command generated the box-plot and the line (yline) that displays the mean for the transformed case filing data. The mean and other statistics for the transformed data were produced by the "summary" command.

```
su LogTF, d
```

N.B. the information displayed in Table 4.1 was also generated using the following STATA command.

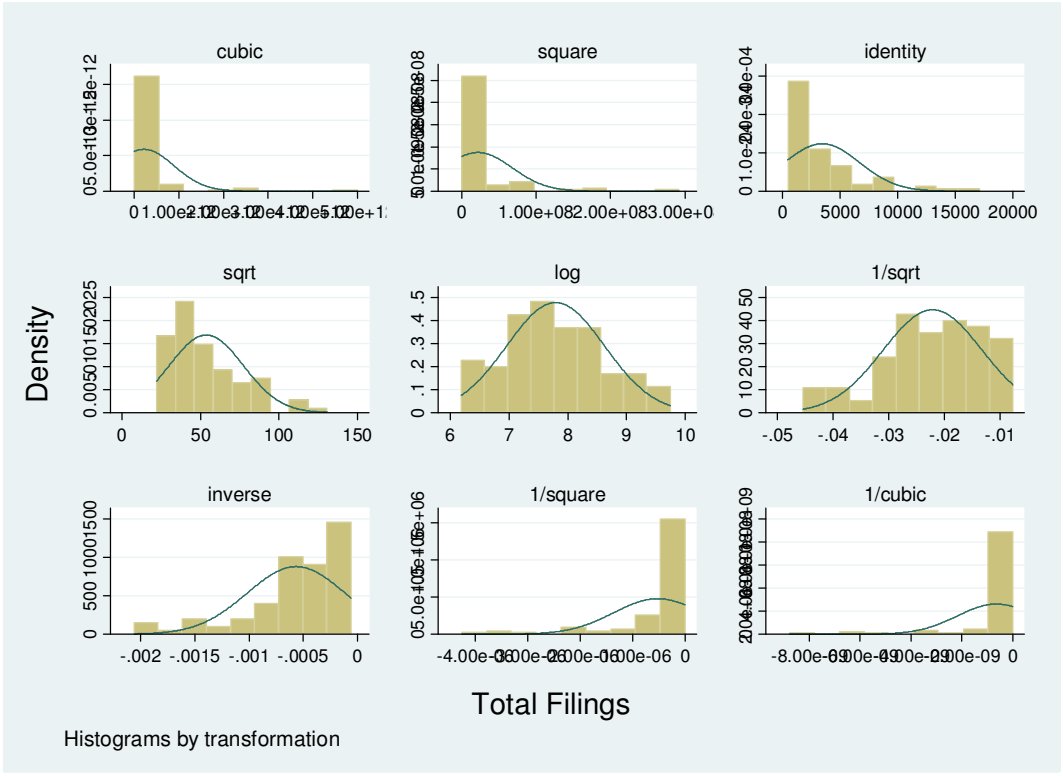
```
list disctrictcourt totalfilings if totalfilings > 8400
```

[The 8400 number was determined by a visual inspection of the stem-leaf of totalfilings. That indicated that the values above 8400 were of interest. This command produced a list of the district courts (disctrictcourt is the name of each district court in the data set) and the case filings for those districts if the value of the case filings exceeded 8400.]

The discussion of transformations indicates that the process is trial and error. There is a STATA command that provides a useful, quick visual assessment of the transformation ladder for a variable. The only trouble with this routine is that the objective and the results focus on making the distribution normal rather trying to make the variable more symmetrical. There is no indication of the symmetry of the transformations, although the visual inspection of the histograms provides a chance to assess symmetry. The routine and the result for the **case filings** data are presented here.

```
gladder totalfilings
```

The result of the gladder command follows. It is clear from this display of histograms that the log transformation is the most successful in converting the variable into a normal distribution. It is also evident that the symmetry of the log transformation is the best approximation to a symmetrical distribution compared with the other transformations presented by this routine.



From this set of histograms one can see several points. First, the "identity" distribution in the upper right-hand corner is the distribution of the actual data. That is hardly symmetrical or normal. The other transformation that was attempted in the discussion in the chapter is labeled the "inverse" here. That transformation hardly produces a satisfactory result here either.

It is not recommended that one rely on this routine to determine which transformation, if any, is the best. The STATA Manual indicates that this routine is "useful pedagogically," but it should not be relied on for research or analysis. This is a quick method of assessing the impact of various transformations on a variable. However, the objective of the routine is to normalize the variable, NOT make it more symmetrical. We strongly recommend trying the transformations. The analyst should look at each transformation, and assess the change in the IR and the OR (the ratios). That provides a much better understanding of the variable than does the gladder routine. In the discussion above regarding transformations, it was suggested that "trying" various transformations moving up or down the transformation ladder would provide a very good sense of the variable and developing symmetry.

Bibliography

Cleveland, William. 1994. **The Elements of Graphing Data** rev. ed. (Hobart Press for AT&T).

Hartwig, Frederick & Brian Dearing (1979). **Exploratory Data Analysis** (Sage Publications.).

Johnson, Janet Buttholf, H.T. Reynolds, and Jason D. Mycoff (2008). **Political Science Research Methods** 6th ed. (Co gressional Quarterly Press).

Kellstadt, Paul M. & Guy D. Whitten (2009). **THE FUNDAMENTAL OF POLITICAL SCIENCE RESEARCH** (Cambridge University Press).

McNeil, Donald (1977). **INTERACTIVE DATA ANALYSIS: A PRACTICAL PRIMER** (Wiley-Interscience).

Mosteller, Frederick & John Tukey, (1977). **DATA ANALYSIS AND REGRESSION: A SECOND COURSE IN STATISTICS** (Addison-Wesley).

Nachmias-Frankfort, Chava and David Nachmias (2007). **Research Methods in the Social Sciences** 7th ed. (Worth Publishers).

Pollock, Philip. (2005) **THE ESSENTIALS OF POLITICAL ANALYSIS** 2d ed. (Congressional Quarterly Press).

Pollock, Philip. (2009) **THE ESSENTIALS OF POLITICAL ANALYSIS** 3rd ed. (Congressional Quarterly Press).

Robbins, Naomi. (2005) **Creating More Effective Graphs** (Wiley Interscience).

Salkind, Neil J. (2006) **Explaining Research** 6th ed. (Pearsion Education Inc.).

Trochim, William M.K. (2001) **The Research Methods Knowledge Base** (Atomic Dog Publishing).

Tukey, John (1977). **Exploratory Data Analysis** (Addison-Wesley).

Velleman, Paul F. and David C. Hoaglin (1981) **Applications, Basics, and Computing of Exploratory Data Analysis** (Duxbury Press).