EDA - Exploratory Data Analysis (06-03)

[These notes are designed to accompany instructor led classes. Major credit is due to the work of the world class statistician, the late John Tukey. This material is inspired by his insights. Rob Rucker, 2008-01-09]

Introduction and Context

Figure out what you *can* do before you try to measure *how well* you have done it! (paraphrase from John Tukey, 1977). My intent here is to set a context for our upcoming quantitative studies, and try to convince you that preliminary work in data exploration is a necessary (and enjoyable) precondition for further data analysis. The field of descriptive statistics and exploratory quantitative analysis, has blossomed into a whole field called Exploratory Data Analysis (EDA) that these notes will introduce to you.

Exploratory Data Analysis has drawn a major following among statisticians as well as engineers and scientists in the last 30 years (for example, all of the graphical techniques of Six Sigma are included within EDA). Starting with John Tukey's seminal work in 1977 (Exploratory Data Analysis (EDA)), right up to the present, there are hundreds of thousands of examples of its real world use.

An analogy might help before we get started looking at ways to explore actual data sets.

Phase I Quantitative: Police & Detectives Collect/Analyze Evidence

Think about our Anglo-Saxon judicial system. This system has several analogies to the approach taken by data-analysts. At the first phase, before there is a trial and a judge, there are evidentiary proceedings carried out by the police and detectives. Their job is to unearth the evidence, and present it to the judge, jury, and lawyers. Theirs' is the *exploratory* phase of jurisprudence and work can't continue until they have fulfilled their job. After all, what these people come up with is all that the material evidence that the judge/jury/lawyers are going to touch, see, and hear. (And in American jurisprudence, this is all the evidence they are allowed to consider)

We are all too aware of the dangers of false, misleading, incomplete, or incorrect evidence, so it makes sense to spend as much time and energy gathering and interpreting evidence as it does to try to judge it in terms of some pre-existent standards. Witness the discovery of new DNA evidence that exonerates a death-row inmate. Obviously, evidence gathering can be the difference between life and death/success or failure. So, the gathering of evidence is deeply important: this phase is much more than merely descriptive (as many textbooks characterize it), it is investigative and indicative!

Phase II Interpretive: Judge & Jury Test Hypotheses and

Make Judgments

After the evidence is collected, there is a second phase, judgement and inferences. The judgements result from presentations and interpretations of evidence that suggest or support various *hypotheses* put forth by the defence and prosecution lawyers. (Note that how the evidence is presented has an enormous influence on what is understood by the audience, hence EDA!). What can be *inferred* from the evidence to support the hypothesis of guilty or not guilty? (Notice that this doesn't say innocent or not innocent). Both sides use the evidence to support or refute various hypotheses leading to the ultimate jury decision. It's interesting to note that the jury is reminded to reach their conclusions "Beyond a Reasonable Doubt". Here you can see another idea that will be carried over into statistics known as a "Confidence Level". In American justice, the assumption (hypothesis) is that the person is NOT guilty. It takes a lot of evidence to refute this hypothesis and the role of the jury is to determine if there is enough evidence to do that. Roughly, the jury will want to be 'convinced' with about 95% confidence, that the person *is* guilty. So, if the evidence for conviction is not iron clad, then there is a 'good' chance that the person will go free.

Although this second phase of jurisprudence, the *inferential or judgemental* phase, gets the headlines, keep in mind that it is the weight of accumulated, forcefully presented, exploratory evidence that ultimately tips the scales of justice, one way or the other.

And Now - Statistics

In statistics it's the same. First Phase: You need to gather your evidence (often in the form of personal experience/calculations/artifacts, observations, questionnaires, interviews, software simulations, or secondary sources). Next, you need to look carefully at that evidence, you need to 'feel' what's its like: what does it 'tell you', where can you get more? Here is the stage where EDA techniques can help you present and assess the evidence in ways that can mean the difference between being *out-sourced* or: landing a promotion, bonus, or even a Nobel prize (yes, you can notice and follow-up on something that important)! Instead of presenting your finding to a jury or the opposing lawyers though, you will usually find yourself in consultation with the relevant *domain experts of your organization* (unless you are the expert). The domain experts are the people who know the most about the area the data comes from, and are in a position to appreciate unusual features, exposed by EDA. This is the detective phase, just as in the judicial context.

Second Phase: The inferential/confirmatory phase starts when the evidence is in, and is now tested against some pre-conceived hypothesis or some inferences are drawn. At this stage, pre-assumed theory usually plays a major part in knowing what is to count as a valid hypothesis test or inference- such as the assumption that the data comes from 'Identically Distributed Normal Distributions'. There are many tests that can be done, but they all make some assumptions about where the data 'comes from'. The purpose of the first phase is to help the researcher narrow down where that data might have come from and so point the way to valid hypothesis or inferential tests on it. At the end of this confirmatory phase you have usually increased your knowledge but have also unearthed more questions that need to be an-

swered!

Exploration and Confirmation/Inferences go back and forth in both the law and in statistics. We will learn about both phases in due course, but first, this note concentrates on the *Exploratory* phase, the EDA phase.

Characteristics of EDA

EDA is actually more of a philosophy than a collection of techniques, although graphical and semi-graphical techniques are a major component of this philosophy. Here are a set of principles that start to describe this area:

- keep in mind that the objective of analysis is *insight*, not numbers (this is a paraphrase from Richard Hamming, a world class data analyst as well).
- use techniques that (forcibly) reveal the underlying structure of the data set. This almost always means bringing out the visual/auditory/textual aspects of the data set in some compelling way.
- look at the data from multiple perspectives, using multiple techniques and approaches.
- identify, as far as possible, important variables.
- look for and examine unusual occurrences, outliers and anomalies (remember, every unanticipated occurrence could be the signal for a critical insight -- or a Nobel Prize). In manufacturing, the detection of an error/flaw/out-of-spec event is cause for celebration, since some process component needs to be re-thought/reworked and subsequently improved.
- always be on guard to test underlying assumptions. To paraphrase the humorist Will Rogers: It ain't what you don't know that causes trouble but, what you know that ain't so!
- determine tentative models that can be further tested.

Most EDA techniques are graphical since the basic idea is to explore, and graphics are a great way to do that. We humans are pattern matching animals, and so searching for patterns within the data, using graphics, is a natural and attractive approach to take. A great series of books on graphical principles and presentation, besides those of Tukey's, are those of Tufte (1983, 1990, 1997). For the analyst in you, check these out. Tufte shows *why* certain graphic patterns work well and in addition, lays out principles on *how* to make your own graphics very effective.

A First Look: Batch Data with Stem & Leaf Analysis

This first section deals with looking at batches of data as a first cut. How to do that? What does it mean to look at a batch, how to get a 'feel' for that data?

Realize that having a mess of data dumped in your lap will be usually be the first thing that happens to you, whether you collect them or they are 'given' to you. It would be a mistake though, to immediately plug them into a statistics package, get a slew of statistical parameters, and assume the data came from one of the 'nice' distributions such as the Bell Curve/Gaussian/Normal).This is a tempting approach, especially if you are pressured by time and circumstances. Resist! Let's do some

preliminary poking about first.

Reducing Detail by Sorting and 'Cutting'

Let me start with simply presenting a batch of numbers and then we'll see what we can do with them. Here are some prices of second-hand cars found in an old newspaper section on used cars: The first car on the list is priced at $250, the next at $150 and so on.(Note: in general, be suspicious of any car lying on its back, with its wheels in the air, costing less than $200).

Raw values (in dollars $)=

{250,150,732,895,695,1623,1492,1066,1693,1166,688,1333,895,1775,895,1895,795}

In this raw form, there is not much more to say, nothing seems to 'jump out', so, we get to work.

Always Sort the Data - A Simple but Effective Pattern Finding Process

So, to do a little more, first put them in *sorted* order. (Sorting is always a good first step)

SortedRawValues=

{150,250,688,695,732,795,895,895,895,1066,1166,1333,1499,1693,1699,1775,1895}

One way to start to grasp what this data set "says", is to narrow down the 'field of view'. As they stand now, the numbers don't seem to mean a whole lot. Let's try one technique that reduces 'clutter' by simply dropping digits, as described below.

A 'Cutting' Approach

For starters, suppose I just cut off the last 2 digits of the sorted list of numbers above, that's right, just drop them, and so am left with the data set below: You, of course, might not want to do such extreme truncation to the data, but there are benefits to less detail as well: the eye travels more easily over uncluttered terrain.

SortedCutBy100(in 100's) = {1,2,6,6,7,7,8,8,8, 10,11, 13, 14, 16, 17, 18}

Can you see some revealed internal patterns? Let me do a preliminary semi-graphical plot of this data set. I am going to count how many times each number appears in the list and replace each number with an 'x', as below. You may recognize this as a *histogram*.

Example S1:

Replacing each number above by an "x", I will plot their frequency of occurrence.

(UNITS=100's of dollars $)

1 | x

2 | x

3 |

4 |

5 |

6 | x x

7 | x x

8 |x x x

9 |

10|x

11|x

12|

13|x

14|x

15|

16|x

17|x

18|x

A First Look at the Data Distribution

So, by just counting "x's" I have what is called a *histogram*. This shows, for example, that I have one "1", two "6's", two "7's", and 3 "8's". Now I can see the *distribution* of the numbers in this data set. Notice the "bunches" of numbers around the $600 to $800 range and the general trailing off to larger values. The idea of a distribution is fundamental in all data analysis and much effort is devoted to describing it. The question is: *where does this data come from*? The payoff is that *if* you can identify the distribution the data comes from (or an approximation), there are usually standard analyses you can do with statistical packages and standard procedures to extract more information (with minimal effort).

Technically, you are looking/hoping to see if your data comes from some cataloged distribution, called *reference* distributions. A few important ones are: the Gaussian (also called the Normal or the Bell Shaped Curve), Binomial, Poisson, Exponential, Geometric, Uniform, and dozens of others. If you can assume your data follows one of these distributions, you are in *good shape*, otherwise, more analysis and more data is desirable.

A Little More Detail Captured - Stem & Leaf Plot

Let me now introduce a more specialized kind of histogram, one that retains a bit more of the original data but doesn't need more writing. This is in keeping with a general principle of graphical presentation: minimize data ink - maximize information (see the Edward Tufte books referenced at the end of this tutorial). That is, make every mark count. Here, instead of an 'x', I replace that "x" with the actual digit it replaced. This provides more information and so is an improvement. This will be called a *stem and leaf plot* (again, this innovation is due to J. Tukey). The structure is that the 'base' part of each number is placed to the left of a vertical line and the remaining digit (or digits) are written to the right. Asterisks are used to 'pad' out the base column so you don't have to write down as much. An example will help:

Example S2 (One Digit Leaf)

Below is shown a Stem and Leaf Plot with a one digit leaf.

Here are the original car price data with the last digit dropped (cut).

SortedCutBy10(in 10's) =

{15,25,68,69,73,79,89,89,89,106,116,133,149,169,169,177,189}

(In the plot below, the first entry represents 15 tens, or $150, the second line represents $250, while the sixth line represents two values, $680 and $690). That sixth line had two leaves, an '8' and a '9'.

[NOTE: When you see an asterisk in the stem area, it is to be (mentally) replaced with the leaf in the leaf area. This is the way you can tell how many digits are to be in the leaves.]

(UNITS = 10's of dollars)

```
1*  |5
2   | 5
3   |
4   |
5*  |
6   | 8 9
7   | 3 9
8   | 9 9 9
9   |
10  |6
11  |6
12  |3
13  |
14  | 9 9
15  |
16  | 9
17  | 7
18  | 9
```

Example S3 (Back to the Original Data Set)

This next plot goes back to all of the original information where the full price is represented. For example, this could be presented to a car dealership manager for study and comment: why are the values bunched up around the $600 to $800 mark? My job as an analyst, or your's, is to bring to the attention of the *domain expert,* features of the data that they might miss without your exploratory analyses.

Stem & leaf with 2 digit leaves, Unit=$1

Sorted = {150,250,688,695,732,795,895,895,895,1066,1166,1333,

1499,1693,1699,1775,1895}

Here I have plotted the actual car prices. Since there are two asterisks in the stem area, that means that the leaves consist of two digits. So, 1**|50 is read as $150 and 2 | 50 is read as $250.

(UNITS= unit dollars)

1**| 50
2 | 50
3 |
4 |
5 |
6 | 88, 95
7 | 32, 95
8 | 95, 95, 95
9**|
10 | 66
11 | 66
13 | 33
14 | 99
15 |
16 | 93, 99
17 | 75
18 | 95

In the above stem & leaf, the first line is interpreted as: "$150" while line 6 has a stem = 6 and leaf = 95, which is interpreted as "$695"

Example S4 (An EDA Type Analysis that DID Lead to a Nobel Prize)

The next example uses data that pointed to an important discovery. The discovery of Argon gas. The researcher in this case, Lord Raleigh, was a distinguished physicist before he made this discovery but this discovery further enhanced his reputation.

Lord Raleigh, 1894. "On an anomaly encountered in determinations of the density of nitrogen gas" as quoted in Tukey 1977. Raleigh analyzed the weights of a standard volume of nitrogen from different researchers [this is a data collection phase from secondary sources]. The kicker here was that some of the sources of analyses were from air and some were from chemical substances. The weights displayed some interesting characteristics that made Raleigh conduct a more careful investigation, resulting in a major discovery. If you plot these results using stem an leaf plots, you too could have deduced the presence of an anomaly. (Raleigh got a Nobel prize for this and related work)

Raw weights, from non-air (chemical compound) sources

{2.30143, 2.29816, 2.30182, 2.29890, 2.29889, 2.29940, 2.29849, 2.29889}

Raw weights from air {2.31017, 2.30986, 2.31010, 2.31001, 2.31024, 2.31030,

2.31028}

The trick here is to figure out what to use for the stems. Try several different ways and see what you discover (you may have to consider stems of 3 or 4 digits!) You should see a separation of points, but why? The separation of weights from the categories of sources led Raleigh to conclude there was another element present in air, later identified as *Argon*.

Lessening the Distractions of Fractions and Negative Numbers

This next example is just to illustrate various labor saving devices as well as 'smoothing' out the data values so that they don't catch the eye and distract from seeing important features. Decimal points and negative numbers are in this 'distracting' category so lets see what we can do with them (or without them).

Smoothing out batches of numbers for an easier first look

Practical arithmetic - rounding (round to nearest number, 0.5 goes to even number)

Rounding 15.5 ->16

Rounding 16.5 ->16

Rounding 16.7 -> 17

Practical arithmetic - cutting

Cutting 15.5 -> 15

Cutting 16.7 -> 16

The presence of the decimal point ".5" occurs often, but it disturbs the visual look of data, so for viewing purposes use an "h" when you need to write it down. For example, if you saw "15.5", consider replacing it by 15h if you need to keep that data value, otherwise just chop off the fraction.

Digits in numbers that only serve as place holders, can be replaced by asterisks "*". In other words, it may not be helpful to report the population of dogs in the Metro Phoenix valley as 856,234 dogs. It might be sufficient for your purposes to only quote 856*** thousand, where the reader is to infer that the remaining 3 digits are between 000 and 999.

Using 2 Line Stems for finer discrimination

An example of a 2-line stem & leaf plot follows. You might want to try this when there are just too many leaves on a single line: break the stem into two lines and use a different symbol for the 'branched' stem.

Let the first stem "*" stand for leaves 0-4, while the second stem "." will represent leaves from 5-9. Note that this evenly divides the numbers into categories of 5 each.

Example S4 - 2 Line Stem & Leaf

Do a 2 line stem&leaf of the data set below:

test4 = {40, 40, 41, 42, 42, 43, 45, 46, 46, 47, 47, 49, 50, 50, 51, 51, 55, 57, 57, 59, 60, 60, 61, 64, 68, 69}

4 * |0 0 1 2 2 3

- |5 6 6 7 7 9

5 * |0 0 1 1

- |5 7 7 9

6 * |0 0 1 4

- |8 9

Using 5 Line Stems for Maximum Discrimination - 5 line Stem&Leaf Plots

The English language seems well set up to help with this categorization of the data. Now I am going to break up the stems into 5 subdivisions (lines) as follows:

* - holds 0, or 1 leaf digit values

t - holds two or three, 2, or 3

f - holds four or five, 4, or 5

s - holds six or seven, 6, or 7

• - holds 8, or 9

Example S5 5-Line Stem & Leaf

Work on the data set above, Example S4, and do a 5 line stem & leaf

4 * |0 0 1

t |2 2 3

f |5

s |6 6 7 7

- |9

5 * |0 0 1 1

t ||

f |5 5

s |7 7

- |9

6 * |0 0 1

t |4

f |

s |

- |8 9

Tallies (Raw Counting)

Perhaps an even more basic operation when first looking at a data set is to count its members! That can't be too hard, right? Well up to this point you may only know one way to do this, the 'slash' method. This is the most common way to tally numbers of entries in a batch. What I am talking about here is the approach shown in the diagram below "Slash versus Block Tally Schemes" on page 10. This approach tends to produce errors when I close off three slashes rather than four. Additionally,

the “slash tally” style encourages multiple forward slashes with no distinction between them.

A better way that prevents more mistakes was introduced by Tukey who credits it to a practice by North American indians. I have named it “block counting” and is made up of ‘dots’ and lines. The merit of this scheme is that it uses ‘dot’s and lines to show cumulative amounts and summarizes blocks of counts in units of 10 and so is “finger friendly”. The block tally lets you use several configurations to represent a given number, but they are consistent. What I mean is that for ‘6’, there are 4 ways you could show it and they are equivalent.Give this a try the next time you need to count up a bunch of numbers. I think you will be pleasantly surprised at its efficiency and ease of checking. (Did you notice the error in the slash tally).

.

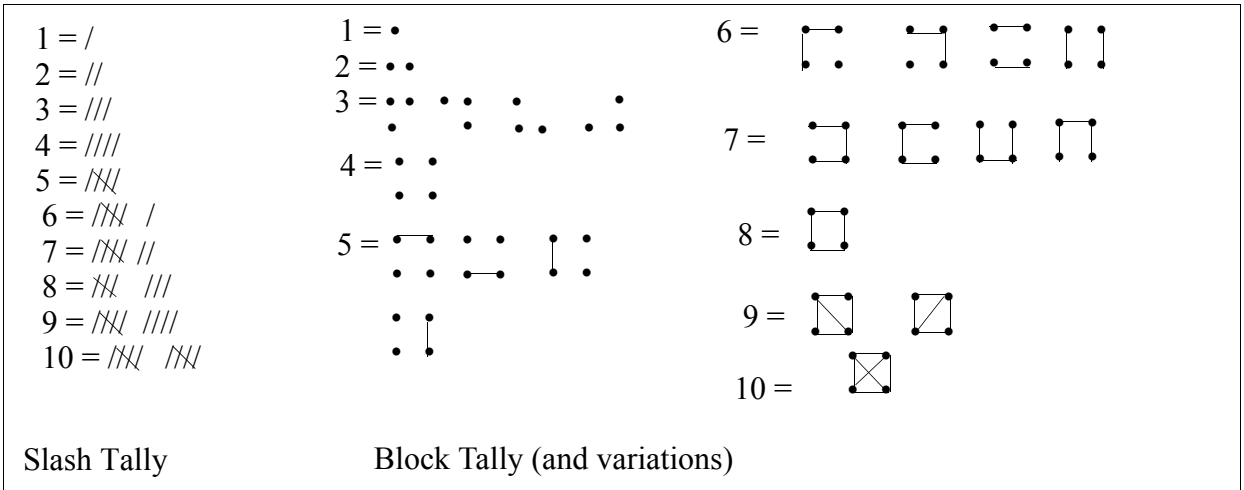

FIGURE 1. Slash versus Block Tally Schemes

Summary

I hope you have developed some feeling for how to start looking at batches of numbers. The intent is to discover if there is something that is not apparent from the mass of values you encounter at first glance. Tukey has provided some help in this area with his innovation of “Stem & Leaf Plots” as well as other graphic models we will encounter subsequently. We will see how they lead into the next topics when we summarize the batch using parameters we get from the Stem& Leaf Plots. The next tutorial in this series, Five Number Summaries & Box Plots, extends the ideas found here leading to a compact semigraphical that tutorial shows how the box and whisker plots may be used in trend analysis.

References

Tukey, John, (1977), *Exploratory Data Analysis*, Wiley, New York

Mosteller, D.& J. Tukey (1977), *A Second Course in Data Analysis*, Wiley, New York.

Tufte, Edward, (1983) *Graphical Display of Quantitative Data*, Graphics Press, Cheshire Connecticut.

Tufte, Edward, (1990) *Envisioning Information*, Graphics Press, Cheshire Connect-

icut

Tufte, Edward, (1997) *Visual Explanations*, Graphics Press, Cheshire Connecticut