Introduction:
What Is Data Analysis?

What is the wealth of the United States? Who's got it? And how is it changing? What are the consequences of an experimental drug? Does it work, or does it not, or does its effect depend on conditions? What is the direction of the stock market? Is there a pattern? What is the historical trend of world climate? Is there evidence of global warming? — This is a diverse lot of questions with a common element: The answers depend, in part, on data. Human beings ask lots of questions and sometimes, particularly in the sciences, facts help. *Data analysis is a body of methods that help to describe facts, detect patterns, develop explanations, and test hypotheses. It is used in all of the sciences. It is used in business, in administration, and in policy.*

The numerical results provided by a data analysis are usually simple: It finds the number that describes a typical value and it finds differences among numbers. Data analysis finds averages, like the average income or the average temperature, and it finds differences like the difference in income from group to group or the differences in average temperature from year to year. Fundamentally, the numerical answers provided by data analysis are that simple.

But data analysis is not *about* numbers — it uses them. Data analysis is about the world, asking, always asking, "How does it work?" And that's where data analysis gets tricky.

For example: Between 1790 and 1990 the population of the United States increased by 245 million people, from 4 million to 249 million people. Those are the facts. But if I were to interpret those numbers and report that the population grew at an average rate of 1.2 million people per year, 245 million people divided by 200 years, the report would be wrong. The facts would be correct and the arithmetic would be correct — 245 million people divided by 200 years is approximately 1.2 million people per year. But the interpretation "grew at an average rate of 1.2 million people per year" would be wrong, dead wrong. The U.S. population did not grow that way, not even approximately

For example: The average number of students per class at my university is 16. That is a fact. It is also a fact that the average number of classmates a student will find in his or her classes is 37. That too is a fact. The numerical results are correct in both cases, both 16 and 37 are correct even though one number is twice the magnitude of the other — no tricks. But the two different numbers respond to two subtly different questions about how the world (my university) works, subtly different questions that lead to large differences in the result.

The tools of the trade for data analysis begin with just two ideas: Writers begin their trade with their A, B, C's. Musicians begin with their scales. Data analysts begin with lines and tables. The first of these two ideas, the straight line, is the kind of thing I can construct on a graph using a pencil and a ruler, the same idea I can represent algebraically by the equation "y = mx + b". So, for example, the line constructed on the graph in Figure 1 expresses a hypothetical relation between education, left to right, and income, bottom to top. It says that a person with no education has an income of $10,000 and that the rest of us have an additional $3,000 for each year of education that is completed (a relation that may or may not be true).

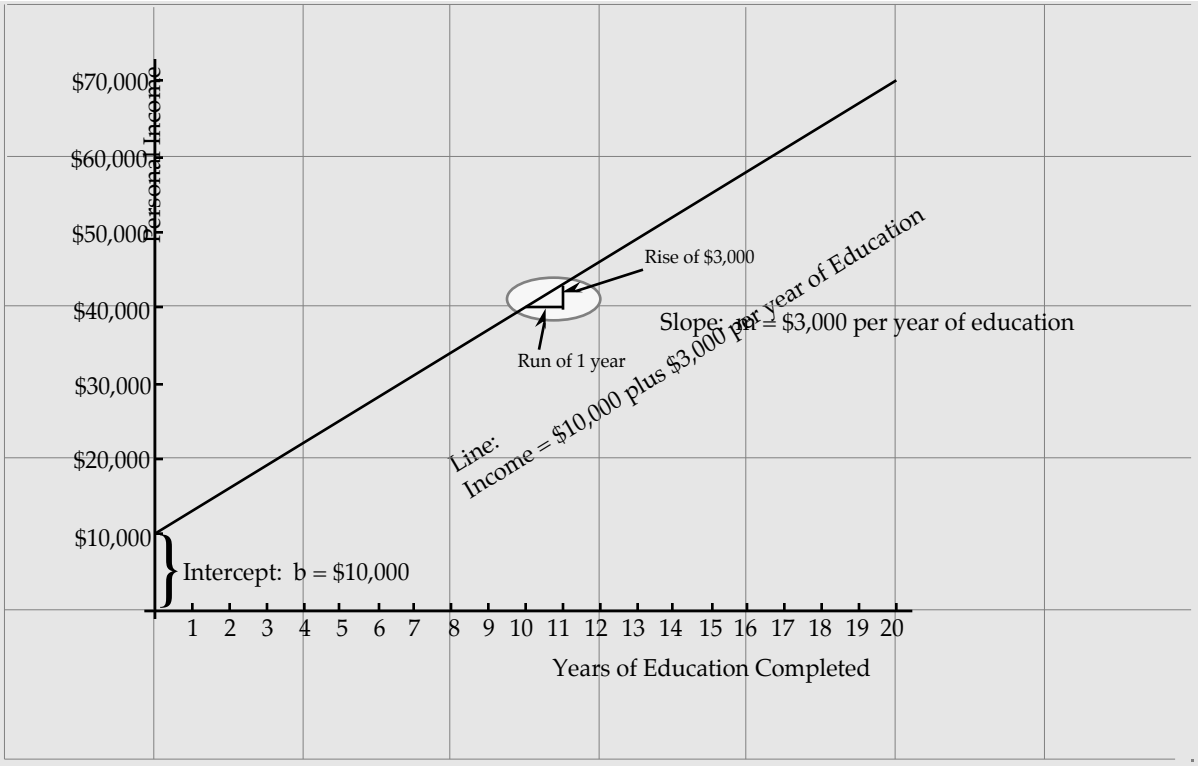

Figure 1
Hypothetical Linear Relation Between Income and Education
The hypothetical line shows an intercept, b, equal to $10,000 and a slope, which is the rise in dollars divided by the run in years, that is equal to $3,0000 per year.

This first idea, the straight line, is the best tool that data analysts have for figuring out how things work. The second idea is the table or, more precisely, the "additive model". The first idea, the line, is reserved for data we can plot on a graph, while this second idea, the additive model, is used for data we organize in tables. For example, the table in Figure 2 represents daily mean temperatures for two cities and two dates: The two rows of the table show mean temperature for the two cities, the two columns show mean temperatures for the two dates.

The additive model analyzes each datum, each of the quantities in the table, into four components — one component applying to the whole table, a second component specific to the row, a third component specific to the column, and a fourth component called a "residual" — a leftover that picks up everything else. In this example the additive model analyzes the temperature in Phoenix in July into

1: 64.5° to establish an average for the whole table, both cities and both dates,

2: plus 7.5° above average for Phoenix, in the first row,

3: plus 21° above average for July, in the second column,

4: plus 1° as a residual to account for the difference between the sum of the first three numbers and the data.

Adding it up,

Observed equals *All Effect* plus *Phoenix Effect* plus *July Effect* plus *Residual* .

That is,

$92^\circ = 64.5^\circ + 21^\circ + 7.5^\circ + 1^\circ$

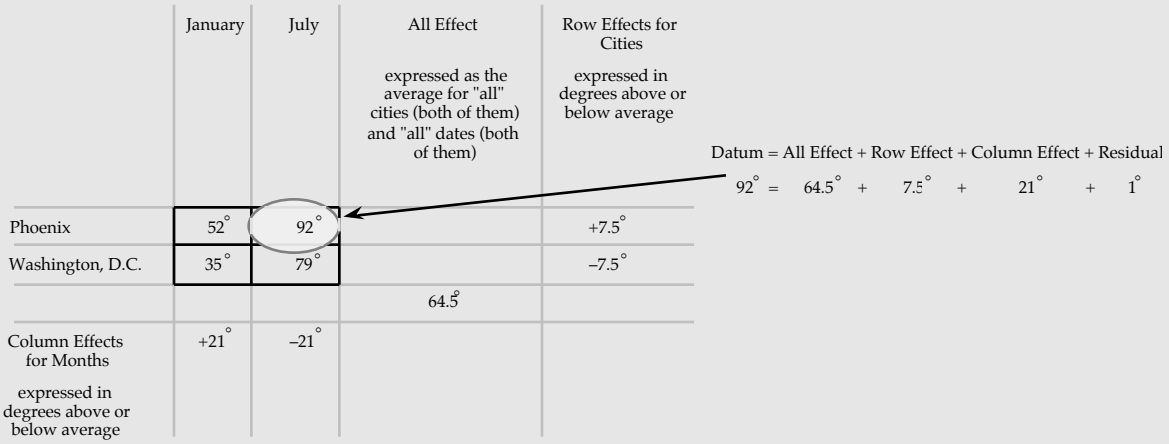

Figure 2

Normal Daily Mean Temperatures in Degrees Fahrenheit

From the Statistical Abstract of the United States, 1987, Table 346, from the original by the U.S. National Oceanic and Atmospheric Administration, Climatography of the United States, No. 81, Sept., 1982. Also note John Tukey's, Exploratory Data Analysis, Addison Wesley, 1970, 0. 333.

There you are, lines and tables: That is data analysis, or at least a good beginning. So what is it that fills up books and fills up the careers of data analysts and statisticians? Things begin to get "interesting", that is to say, problematical, because even the best-behaved data show variance: Measure a twenty gram weight on a scale, measure it 100 times, and you will get a variety of answers — same weight, same scale, but different answers. Find out the incomes of people who have completed college and you will get a variety of answers. Look at the temperatures in Phoenix in July, and you will get a variety, day to day, season to season, and year to year. Variation forces us to employ considerable care in the use of the linear model and the additive model.

And life gets worse — or more interesting: Truth is that lots of things just are not linear: Adding one more year of elementary school, increasing a person's years of education from five to six, doesn't really have the same impact on income as adding one more year of college, increasing a person's years of education from fifteen to sixteen — while completing a college degree. So the number of dollars gained for each extra year of education, is not constant — which means that, often, the linear model doesn't work in its simplest form, not even when you allow for variation. And with tables of numbers, the additive model doesn't always add up to something that is useful.

So what do we do with a difficult problem? This may be the single most important thing we teach in data analysis: Common sense would tell you that what you tackle a difficult problem with a difficult technique. Common sense would also tell you that the best data analyst is the one with the largest collection of difficult "high powered" techniques. But common sense is wrong on both points: In data analysis the real "trick" is to *simplify the problem* and the best data analyst is the one who gets the job done, and done well, with the most simple methods.

Data analysts do not build more complicated techniques for more complicated problems — not if we can help it. For example, what would we do with the numbers graphed in Figure 3? Here the numbers double at each step, doubling from 1, to 2, to 4, to 8, which is certainly not the pattern of a straight line. In this example the trick is

to simplify the problem by using logarithms or the logarithmic graph paper shown in Figure 4 so that, now, we can get the job done with simple methods. Now, on this new graph, the progression, 1, 2, 4, 8,… is a straight line.

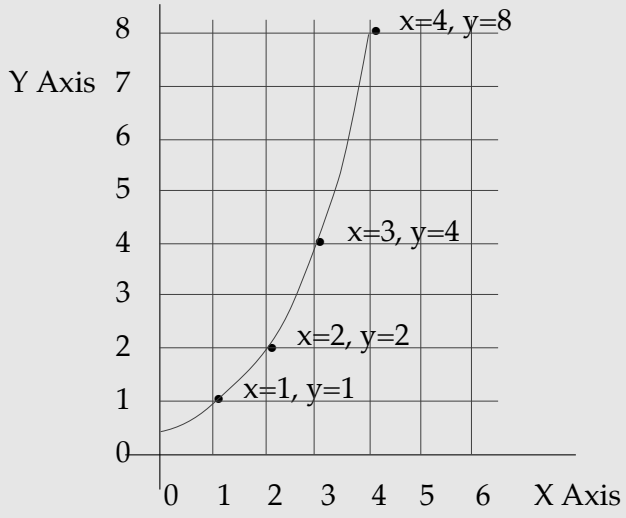

Figure 3
Non-Linear Relation Between X and Y

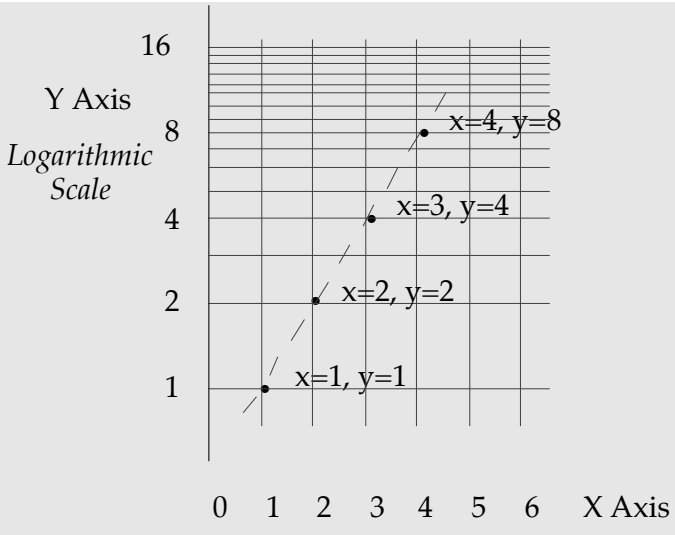

Figure 4
Non-Linear Exponential Relation Between X and Y Made Linear Using a Semi-Logarithmic Graph

"Tricks" like this enormously extend the range of things that an experienced data analyst can analyze while staying with the basics of lines and tables. In sociology, which is my field, this means learning to use things like "log people". In business and economics it means learning to use things like "log dollars". In biology it means learning to use things like the square root of the number of beasties in a drop of pond water or the cube root of the weight of an organism. Learning what these things mean is perhaps the most time consuming part of an introduction to data analysis. And the payoff is that these techniques extend the ability of simple tools, of the line and the table, to make sense of a complicated world.

And what are the *Rules* of data analysis? Some of the rules are clear and easy to state, but these are rather like the clear and easy rules of writing: Very specific and not very helpful — the equivalent of reminders to dot your "i's" and cross your "t's". The real rules, the important ones, exist but there is no list — only broad strategies with respect to which the tactics must be improvised. Nevertheless it is possible to at least name some of these "rules." I'll try the list from different angles. So:

1. Look At the Data / Think About the Data / Think About the Problem / Ask what it is you Want to Know

Think about the data. Think about the problem. Think about what it is you are trying to discover. That would seem obvious, "Think." But, trust me, it is the most important step and often omitted as if, somehow, human intervention in the processes of science were a threat to its objectivity and to the solidity of the science. But, no, thinking is required: You have to interpret evidence in terms of your experience. You have to evaluate data in terms of your prior expectations (and you had better *have* some expectations). You have to think about data in terms of concepts and theories, even though the concepts and theories may turn out to be wrong.

2. Estimate the Central Tendency of the Data.

The "central tendency" can be something as simple as an average: *The average weight of these people is 150 pounds.* Or it can be something more complicated like a rate: *The rate of growth of the population is two percent per annum.* Or it can be something sophisticated, something based on a theory: *The orbit of this planet is an ellipse.* And why would you have thought to estimate something as specific as a rate of growth or the trace of an ellipse? Because you thought about the data, about the problem, and about where you were going (Rule 1).

3. Look at the Exceptions to the Central Tendency

If you've measured a median, look at the exceptions that lie above and below the median. If you've estimated a rate, look at the data that are *not* described by the rate. The point is that there is always, or almost always, variation: You may have measured the average but, almost always, some of the cases are not average. You may have measured a rate of change but, almost always, some numbers are large compared to the average rate, some are small. And these exceptions are not usually just the result of embarrassingly human error or regrettable sloppiness: On the contrary, often the exceptions contain information about the process that generated the data. And sometimes they tell you that the original idea (to which the variations are the exception) is wrong, or in need of refinement. So, look at the exceptions which, as you can see, brings us back to rule 1, except that this time the data we look at are the exceptions.

That circle of three rules describes one of the constant practices of analysis, cycling between the central tendencies and the exceptions as you revise the ideas that are guiding your analysis. Trying to describe the Rules from another angle, another theme that organizes the rules of evidence can be introduced by three key words: falsifiability, validity, and parsimony.

1. Falsifiability

Falsifiability requires that there be some sort of evidence which, had it been found, your conclusions would have had to be judged false. Even though it's your theory and your evidence, it's up to you to go the additional step and formulate your ideas so they can be tested — and falsified if they are false. More, you yourself have to look for the counter evidence. This is another way to describe one of the previous rules which was "Look at the Exceptions".

2. Validity

Validity in the scientific sense, requires that conclusions be more than computationally correct. Conclusions must also be "sensible" and true statements about the world: For example, I noted earlier that it would be wrong to report that the population of the United States had grown at an average rate of 1.2 million people per year. — Wrong, even though the population grew by 245 million people over an interval of 200 years. Wrong even though 245 divided by 200 is (approximately) 1.2. Wrong because it is neither sensible nor true that the American population of 4 million people in the United States in 1790 could have increased to 5.1 million people in just twelve months. That would have been a thirty percent increase in one year — which is not likely (and didn't happen). It would be closer to the truth, more valid, to describe the annual growth using a percentage, stating that the population increased by an average of 2 *percent* per year — 2 *percent* per year when the population was 4 million (as it was in 1790), 2 *percent* per year when the population was 250 million (as it was in 1990). That's better.

3. Parsimony

Parsimony is the analyst's version of the phrase "Keep It Simple." It means getting the job done with the simplest tools, provided that they work. In military terms you might think about weapons that provide the maximum "bang for the buck". In the sciences our "weapons" are ideas and we favor simple ideas with maximum effect. This means that when we choose among equations that predict something or use them to describe facts, we choose the simplest equation that will do the job. When we construct explanations or theories we choose the most general principles that can explain the detail of particular events. That's why sociologists are attracted to broad concepts like social class and why economists are attracted to theories of rational individual behavior — except that a simple explanation is no explanation at all unless it is also falsifiable and valid.

I will be specific about the more easily specified rules of data analysis. But make no mistake, it is these broad and not-well-specified principles that generate the specific rules we follow: Think about the data. Look for the central tendency. Look for the variation. Strive for falsifiability, validity, and parsimony. Perhaps the most powerful rule is the first one, "Think". The data are telling us something about the real world, but what? Think about the world behind the numbers and let good sense and reason guide the analysis.

Reading:

Stephen D. Berkowitz, *Introduction to Structural Analysis*, Chapter 1, "What is Structural Analysis," Butterworths, Toronto, 1982; revised edition forthcoming, Westview, Denver, circa 1997.

Stephen J. Gould, "The Median Isn't the Message," *Discover*, June, 1985.

Charles S. Peirce, "The Fixation of Belief", reprinted in Bronstein, Krikorian, and Wiener, *The Basic Problems of Philosophy*, 1955, Prentice Hall, pp. 40- 50. Original, *Popular Science Monthly*, 1877.