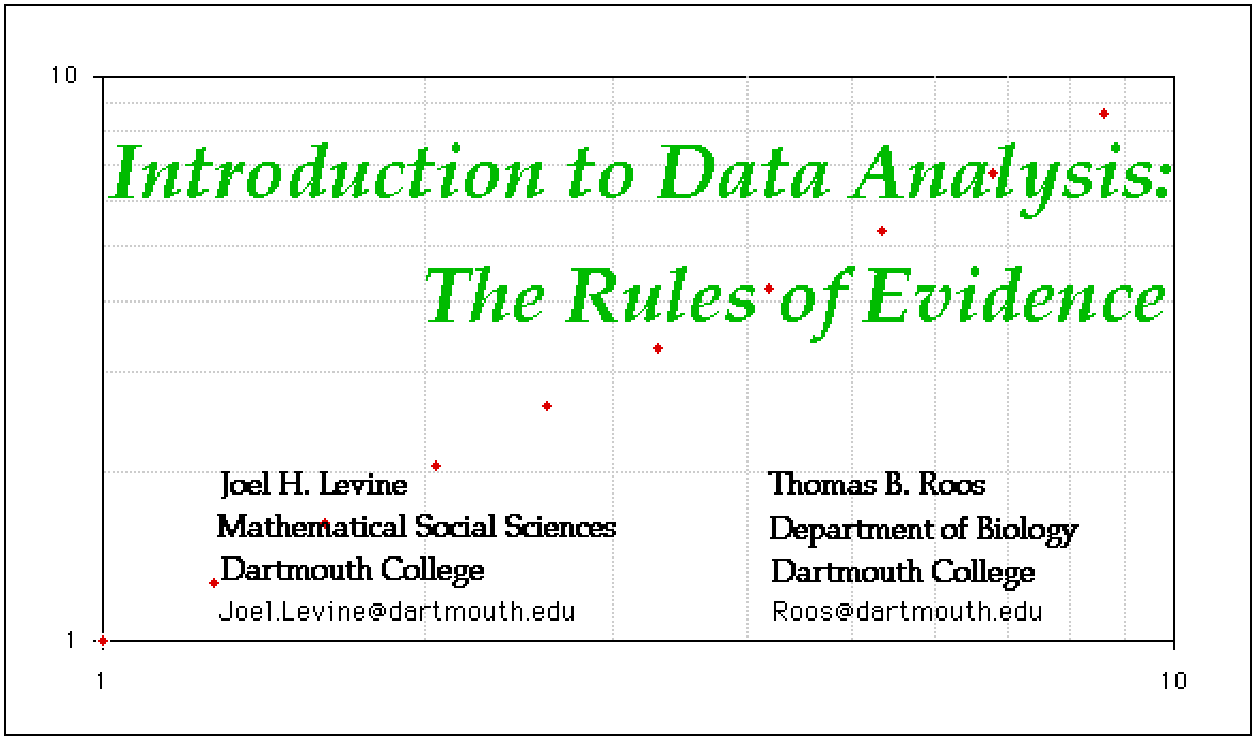


INTRODUCTION TO DATA ANALYSIS: THE RULES OF EVIDENCE

Edited by JOEL H. LEVINE and THOMAS B. ROOS

VOLUME I: WELL-BEHAVED VARIABLES COPYRIGHT (C) 1994-2002, JOEL H. LEVINE, ALL RIGHTS RESERVED
VOLUME II: LINEAR RELATIONS COPYRIGHT (C) 1994-2002, JOEL H. LEVINE, ALL RIGHTS RESERVED
Forthcoming
VOLUME III: TESTS AND ESTIMATES
VOLUME IV: CATEGORICAL DATA

Part of this work has been developed in association with the Math Across The Curriculum Project at Dartmouth College operating under National Science Foundation Support -- Dorothy Wallace, Principal Investigator, Thèrese Stukel, Principal for Statistics and Data Analysis

Address correspondence to **Joel.Levine@dartmouth.edu** and **Roos@dartmouth.edu**. We will be pleased to add links to other Web sites that will be useful to our own students and colleagues and to other users of this site.

Continue to Table of Contents

For regular use of the Web edition of *Rules of Evidence,* bookmark the Table of Contents. In the Web edition of *Rules*, most items referred to in the Table of Contents are ".pdf" files. ".pdf" files are accessible either directly (using Adobe's Acrobat Reader) or indirectly (using a Web browser that uses Adobe's Acrobat Reader). You (or your browser) must have Adobe Acrobat Reader (or have it in your browser's helpers folder. You need it. It is free, but you need it. Ask your local computer guru. The exceptions (the non pdf files) are, or will be, "stuff" like MacIntosh Excel files or straight text files. If you need to get a copy of Adobe Acrobat Reader, begin your odyssey at http://www.adobe.com

INTRODUCTION TO DATA ANALYSIS: RULES OF EVIDENCE

VOLUME I: WELL-BEHAVED VARIABLES

(One-Variable Analysis)

Table of Contents

INTRODUCTION TO DATA ANALYSIS: RULES OF EVIDENCE

VOLUME II: LINES

Introduction to Data Analysis: Rules of Evidence

Preface

Every professor wants every student to believe that the subject being taught is exciting, beautiful, and the most worthy calling a human being could possibly follow. And I of course, want students to feel that way about data analysis. But, in a sense, I have to insist — because if you approach data analysis mechanically, as a matter of ritual and routine, then you can not do it well. A computer, or a human acting like one, can not be programmed to use insight, to have hunches, good or bad, and to track the hunches down. And these things are absolutely essential to data analysis, every bit as necessary as solid logic and precision. Data analysis requires logic and clear thinking, but also style, intuition, eccentricity, inspired guesses, and hot pursuit of good ideas — that may or may not turn out to be correct.

Data analysis is, in part, a body of techniques for dealing with numbers: The numbers can be averages — the average income of a population. The numbers can be counts — the number of protozoa in a drop of pond water. The numbers can be temperatures or velocities — there is no end to the list. But there is a defining characteristic that separates data analysis from mathematics: In data analysis the numbers carry a message from the real world to us. And we look to the data in order to figure out how that world works.

That's what's special about data analysis: Put "two of this" and "two of that" together in a test tube and — reality governs: You'd better check to see just what it is you've got. Granted that "two plus two" will always be "four," but that is a statement in mathematics where "two" and "plus" and "four" are reasonably well-defined abstractions dwelling in the human brain. By contrast, in data analysis the "input" is from the real world and the "output" is tested against the real world. However compelling the assumptions, however logical the conclusion, however precise the mathematics and convincing the theory, the results must be tested against the real world. It is arguable that there are more general kinds of data, data that are not numerical. But, in *Rules*, we, the co-authors, are going to talk about the analysis of data, numerical data, and basic strategies for deciphering the message that they carry.

Preface to Faculty

I propose a test and a question: Collect the data that describe the gross national product of the United States, in sequence, for the last one hundred years. Plot the data on graph paper. And then ask someone to interpret it. That's the test. And the question is — How many people will be able to interpret it? I am not sure of the answer to that question, but I know that most of my colleagues — who teach undergraduates in both the natural and social sciences — are none too sanguine about the probable results for their own students.

And I submit that this is a serious problem. If I were to state that the gross domestic product of the United States had increased an average of 3% per year during the last 100 years, most listeners would claim to understood what I had said. If I were to state that the population of the United States had increased by 2% per year, during the last 200 years, most people would claim to have understood the meaning of the statement. But they do not: If an educated person can not compute such an average, 3% per year, or 2% per year, directly from the data then, in truth, that person does not understand the average — not what the average means, not really. And that's not good — not for the education of scientists, not as preparation for business, not for policy, nor for education in the liberal arts.

I'm afraid to speculate what this means for the general public. I know it means that when I stand up to argue about tax *rates* in the forum of my town meeting, I'm kidding myself. I know it means that public debate about the rate of increase of U.S. health care costs vis à vis the *rate* of inflation — whatever that debate is about — will not be won, lost or, perhaps, even influenced very much by facts, not in the public domain. And, for education, the unhappy truth is that the problem is not going to be solved by pabulumizing the curriculum. The worlds we try to understand and control through data, the worlds, of science, and policy, and business, are not going to simplify themselves in order to accommodate the deficiencies of the analyst. Rather, we are going to have to bring our students up to the demands of the data. To do otherwise is not only wrong but an act of hubris. If, for example, the relation between a person's years of education and a person's income is not linear, then it does no good to describe it as linear anyway "because" linear analysis is easier. That's not an approximation or simplification. That's wrong. And, very likely, all the basic statistical machinery called upon to support that description and lend to it an air of verisimilitude — invoking Gaussian probabilities, confidence intervals, significance tests, and so forth — will not set right what was wrong to begin with.

It's easy enough to offer broad generalities to explain-away the difficulties encountered by students — something about the decline of Western Civilization, or the decay of the scientific subculture, or, at the least, the failure of our educational systems. But these generalizations do not hold up. The problem and the solution lie elsewhere: In college we teach the elite, not the average. We speak to students who have made the cut of standardized tests. And we, the faculty, are not asking these students for technical skills they do not have: They all know what a straight line is. They all know how to draw a graph. Almost every undergraduate comes equipped with some knowledge of algebra. They have had an encounter with logarithms and most, or many, of our students have had an introduction to the calculus.

And, in truth, the mathematics and interpretative skills we, the faculty, have come to avoid in introductory data analysis and statistics classes are technically simpler than the mathematics we teach: For example, logarithmic relations and their interpretation, which we usually omit from data analysis, are technically simpler than the basics of probability, Gaussian distributions, significance tests, Chi-squares and F-ratios which we teach.

So what's wrong? The problem is that students have trouble using the tools they have acquired, carrying their abstract knowledge across the line from abstraction to use: The problem is connecting the math they know to what our social science colleagues call the "substance" behind the methods, to what our physical science colleagues call the "physical intuition" — connecting the math to the data.

I offer these comments to prepare you for what follows: *Rules* is neither more nor less difficult than standard approaches to "methods." It is different. It is based on a different diagnosis of the problems of teaching and learning data analysis. If I am right, or to the extent that I am right, faculties can not solve our students' problem with data analysis by sending them back to the mathematics department for additional preparation in mathematics. I certainly encourage additional training in mathematics, as much as the students can get, but it's not going to solve the problem of connecting abstractions to reality. That's the problem and the way to solve it is to lead the students back and forth between the two cultures, between math and data, between equations and interpretations, passing back and forth, repeatedly and redundantly, until the path is well worn. That is something we in the sciences must do for our own students — it is not the business of mathematicians.

Plan of the Course

Rules of Evidence follows a straightforward outline, beginning , in Volume I, with the study of single variables and advancing, in Volume II, to two variables. With one variable it introduces the use of distributions of the data and the use of summary statistics for the center and the variation. With two variables it introduces the concept of correlation between variables and linear relations including regression.

Interpretation

Within the outline, one distinction of *Rules* is its emphasis on interpretation, in English, accompanied by visual and graphical displays. It is easy for a student to become absorbed in technical details — harder to connect the data to reality. But data analysis is without purpose if it does not connect to reality.

In data analysis technical errors — dropping a decimal place, or forgetting a square root — are not mere technical errors of no consequence. An error usually leads to an absurd interpretation — to estimated growth rates that are ridiculous, to predictions of wealth that are impossible, to drug treatments that kill — and if the analyst does not see the absurdity introduced by "technical error" it means that the connection between the numbers and the data has been lost.

For this reason no example presented to students, and no homework presented by students, should be considered complete without an intelligible write up, in English: A number, like "3", is not the answer to any exercise in data analysis. No matter how long the file of computer output, no matter how pretty the graphs — the answer has to make sense. "3 dollars," or "3 pounds of potatoes per pound of fertilizer," or "3% increase in population per year," or "a dose of 3 grams of antidote per kilogram of body weight" — may be an appropriate statement about data.

Homework

This has an important implication for homework: For my own classes I assign nightly homework but, many decades later, I have accepted the prissy doctrine of grade-school English teachers to the effect that "If you can't put it in writing, then you don't understand it." And to "put it in writing," to produce an intelligent interpretation of data rendered in clear technical English takes time. It is not clear why good technical writing is difficult and time consuming. But, empirically, it is. This means that I assign very few problems — usually no more than one per night. This certainly makes the course easier to teach but, more important, it is the right approach: If I were to assign half a dozen interesting problems for one night's homework, the magnitude of the assignment would, by itself, tell students the opposite of what I intend. The size of the load would tell students to get the numbers right and ignore meaning because it is basically impossible to work out five or six examples and write them up in intelligent English — all in one night.

Exploratory Analysis

Rules of Evidence is heavily influenced by John Tukey's introductory text *Exploratory Data Analysis* in which there is no mention, nor any need to mention probability or Gaussian distributions or, for that matter, least squares or even means and standard deviations — an introduction that Tukey accomplishes without sacrifice of either rigor or precision.

Rules will not go that far, but one reward of reading Tukey's text is to discover, or re-discover, how much can be accomplished well, how much can be accomplished with both rigor and precision — if the analyst has a firm and clear understanding of the basics. When a student can "eyeball" numerical estimates of the center and the deviation, when a student can "read" estimates of slope and goodness of fit right off of a well formed graph, when a student can interpret these things in terms of income, or education, or time, or temperature — as appropriate to the data — then the student is ready for the technology of "r-squared" and standard errors of estimate.

There is a line of thought that says you can not learn the meaning of a thing unless you've got two of them: It forces you to abstract from the examples to the principle. For this reason, I have tried to make it a rule that every time I demonstrate one technical solution to a problem in data analysis I also demonstrate a second: A mean is one realization of the center of a distribution, the median is another. For this reason too, I have tried to introduce data analysis with one broad interdisciplinary course — to be followed by courses that adopt the special practices of the separate disciplines: If the student sees only one solution to a problem, as is usually the case within a mature scientific discipline, it is hard to see that there are others, and then to understand the choices that have been made and the trade-offs — choices that make one solution good or better than another.

Prerequisites

I assume a background in mathematics and, equally important, a willingness to use that background. I do not assume that the student is at ease with that background when it comes to using it, in the real world. That is the business of this course. For example, I use logarithms. I know full well that students are in varying degree "uncomfortable" with such things. Even those who are perfectly comfortable with the "math" need to learn to work with these logarithms when they apply to the size of real world populations, "log people", or when they apply to real world wealth "log dollars". I expect students to work with what they have already learned in secondary school mathematics, and to learn to use it with data. Similarly, I make some (optional) use of calculus, specifically the use of derivatives to find a minimum or a maximum. And, again, I know full well that there will be some discomfort. But the reward is to see how important problems are solved, using the derivative, and to see how calculus invests coherence and strategy to what is otherwise no more than a collection of numerical techniques.

These are reasonable demands of the "average" student in an above average university. It is difficult for faculty to believe this because the "average" student of this generation has a stronger math background than the "average" student of our own (the faculty's) generation. But it is true. If anything, our students may have too much *faith* in mathematics. They have acquired the lay person's idea that mathematics *is* science and they have learned, somewhere, that the more mathematical something looks, the more scientific it is. For a lay person, knowing little of either science or mathematics, that may be a reasonable approximation. For a scientist it is a fatal error.

Technology in The Liberal Arts Curriculum

The reader will see clear traces of my own teaching environment embodied in this text: I teach at an institution that sees itself as a liberal arts college. I do not interpret liberal arts to mean "artsy", and I certainly do not interpret liberal arts to mean "non technical". I interpret it to mean that, at least initially, a liberal arts education attempts to do something broader than teach students a trade. I cannot, it would violate the "rules" of a liberal arts education for me to say to my students, "Here are the symbols and rituals of your trade — learn them!" That's not allowed. The use of the tools and the meaning of the tools go together.

Moreover, data analysis occupies an important niche in the culture. Data analysis is part of the scientific method: Humans have many ways, often strange, for establishing truth. We may rely on authority. We may believe what "everyone" knows to be true. We may find answers in culture and ideology. We may use pure logic. By contrast, science offers, first, skepticism, and then a way to put questions to reality. Data analysis is applied epistemology wherein the questions "What do we know?" and "How do we know it?" become inescapable and must be answered. In later life the skills of the scientist, the business analyst, the policy maker become specialized and differentiated. But in school, at the undergraduate level where these skills begin, they have a common root in the skepticism and the methods of inquiry summarized by the phrase "scientific method".

People like ourselves created these methods to begin with. And people like ourselves have to understand what the methods are for, and how they were invented, and how they may be reinvented or modified when that is what's called for. That's why technical work, including data analysis, belongs in the liberal arts curriculum.

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).

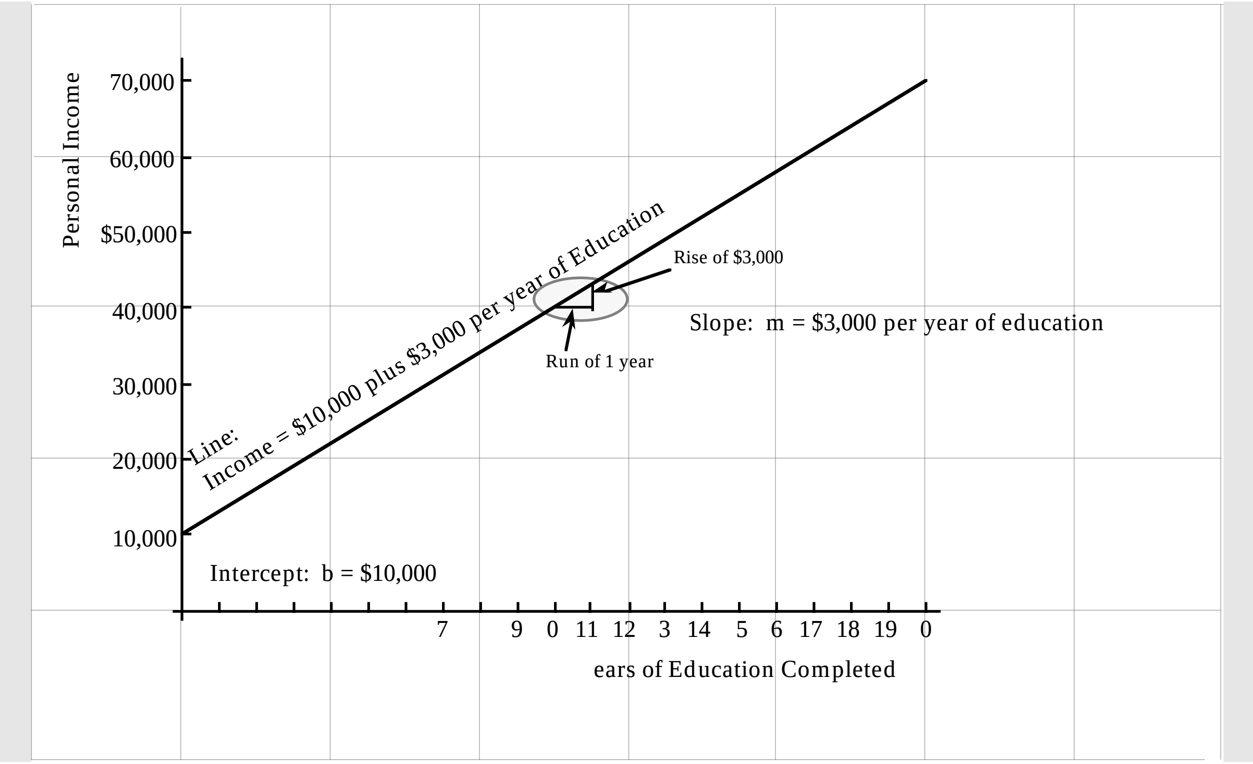


Figure 1
Hypothetical Linear Relation Between Income and Education
The hypothetical line shows an intercept, b, equal to $10,000 and a slope, which is the rise in dollars divided by the run in years, that is equal to $3,0000 per year.

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,

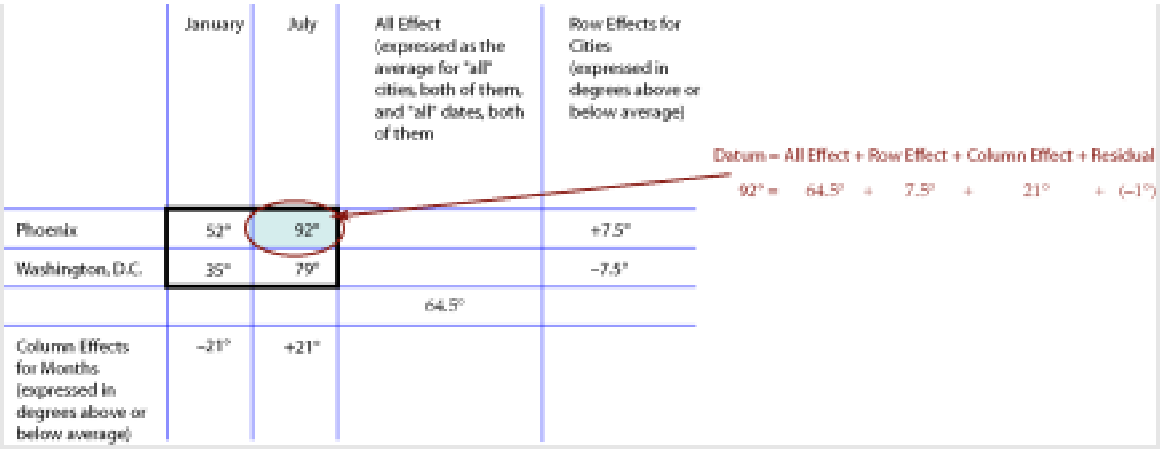
4: plus 1° as a residual to account for the difference between the sum of the first three numbers and the data.

Adding it up,

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

That is,

$$92^\circ = 64.5^\circ + 21^\circ + 7.5^\circ + (-1^\circ)$$



	January	July	All Effect (expressed as the average for "all" cities, both of them, and "all" dates, both of them	Row Effects for Cities (expressed in degrees above or below average)
Phoenix	52°	92°		+7.5°
Washington, D.C.	35°	79°		−7.5°
			64.5°	
Column Effects for Months (expressed in degrees above or below average)	−21°	+21°		

Datum = All Effect + Row Effect + Column Effect + Residual

92° = 64.5° + 7.5° + 21° + (−1°)

Figure 2

Normal Daily Mean Temperatures in Degrees Fahrenheit

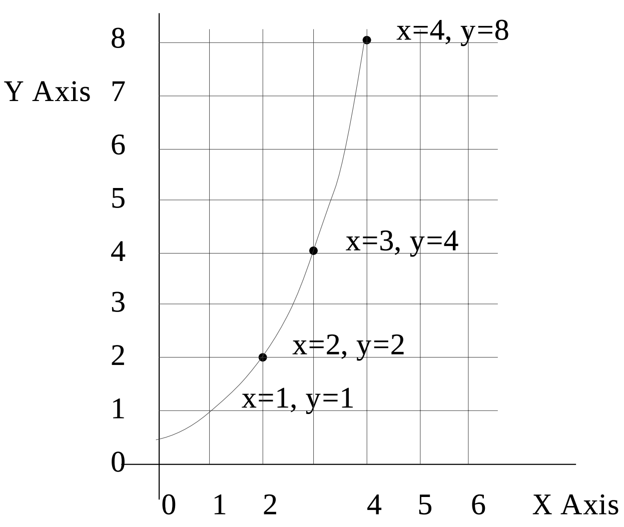
From the Statistical Abstract of the United States, 1987, Table 346, from the original by the U.S. National Oceanic and Atmospheric Administration, Climatography of the United States, No. 81, Sept., 1982. Also note John Tukey's, Exploratory Data Analysis, Addison Wesley, 1970, 0. 333.

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.



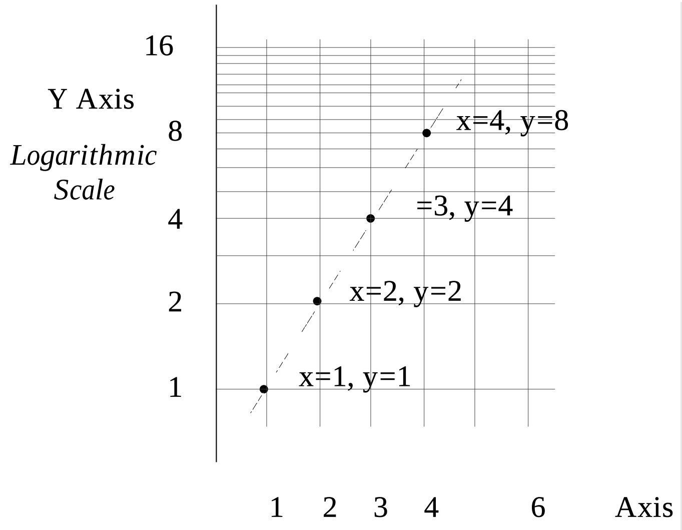


Figure 3
Non-Linear Relation Between X and Y

Figure 4
inear 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.

ESSAY STEPHEN JAY GOULD (From the original source: *DISCOVER*, JUNE 1985.)

The Median Isn't the Message

In 1982,1 learned I was suffering from a rare and serious cancer. After surgery, I asked my doctor what the best technical literature on the cancer was. She told me, with a touch of diplomacy, that there was nothing really worth reading. I soon realized why she had offered that humane advice: my cancer is incurable, with a median mortality of eight months after discovery.

Stephen Jay Gould teaches biology, geology, and the history of science at Harvard.

My life has recently intersected, in a most personal way, two of Mark Twain's famous quips. One I shall defer to the end of this essay. The other (sometimes attributed to Disraeli), identifies three species of mendacity, each worse than the one before, lies, damned lies, and statistics.

Consider the standard example of stretching truth with numbers—a case quite relevant to my story. Statistics recognizes different measures of an "average," or central tendency. The *mean* is our usual concept of an overall average—add up the items and divide them by the number of sharers (100 candy bars collected for five kids next Halloween will yield 20 for each in a just world). The *median,* a different measure of central tendency, is the halfway point. If I line up five kids by height, the median child is shorter than two and taller than the other two (who might have trouble getting their mean share of the candy). A politician in power might say with pride, "The mean income of our citizens is $15,000 per year." The leader of the opposition might retort, "But half our citizens make less than $10,000 per year." Both are right, but neither cites a statistic with impassive objectivity. The first invokes a mean, the second a median. (Means are higher than medians in such cases because one millionaire may outweigh hundreds of poor people in setting a mean; but he can balance only one mendicant in calculating a median).

The larger issue that creates a common distrust or contempt for statistics is more troubling. Many people make an unfortunate and invalid separation between heart and mind, or feeling and intellect. In some contemporary traditions, abetted by attitudes stereotypically centered upon Southern California, feelings are exalted as more "real" and the only proper basis for action—if it feels good, do it—while intellect gets short shrift as a hang-up of outmoded elitism. Statistics, in this absurd dichotomy, often become the symbol of the enemy. As Hilaire Belloc wrote, "Statistics are the triumph of the quantitative method, and the quantitative method is the victory of sterility and death."

This is a personal story of statistics, properly interpreted, as profoundly nurturant and life-giving. It declares holy war on the downgrading of intellect by telling a small story about the utility of dry, academic knowledge about science. Heart and head are focal points of one body, one personality.

In July 1982,1 learned that I was suffering from abdominal mesothelioma, a rare and serious cancer usually associated with exposure to asbestos. When I revived after surgery, I asked my first question of my doctor and chemotherapist: "What is the best technical literature about mesothelioma?" She replied, with a touch of diplomacy (the only departure she has ever made from direct frankness), that the medical-literature contained nothing really worth reading.

Of course, trying to keep an intellectual away from literature works about as well as recommending chastity to *Homo sapiens,*

Discover, June 1985, pp. 42-45.

the sexiest primate of all. As soon as I could walk, I made a beeline for Harvard's Countway medical library and punched mesothelioma into the computer's bibliographic search program. An hour later, surrounded by the latest literature on abdominal mesothelioma, I realized with a gulp why my doctor had offered that humane advice. The literature couldn't have been more brutally clear: mesothelioma is incurable, with a median mortality of only eight months after discovery. I sat stunned for about fifteen minutes, then smiled and said to myself: so that's why they didn't give me anything to read. Then my mind started to work again, thank goodness

If a little learning could ever be a dangerous thing, I had encountered a classic example. Attitude clearly matters in fighting cancer. We don't know why (from my old-style materialisticperspective, I suspect that mental states feed back upon the immune system). But match people with the same cancer for age, class, health, socioeconomic status, and, in general, those with positive attitudes, with a strong will and purpose for living, with commitment to struggle, with an active response to aiding their own treatment and not just a passive acceptance of anything doctors say, tend to live longer. A few months later I asked Sir Peter Medawar, my personal scientific guru and a Nobelist in immunology, what the best prescription for success against cancer might be. "A sanguine personality," he replied. Fortunately (since one can't reconstruct oneself at short notice and for a definite purpose), I am, if anything, even-tempered and confident in just this manner.

Hence the dilemma for humane doctors: : since attitude matters so critically, should such a sombre conclusion be advertised, especially since few people have sufficient understanding of statistics to evaluate what the statements really mean? From years of experience with the small-scale evolution of Bahamian land snails treated quantitatively, I have developed this technical knowledge—and I am convinced that it played a major role in saving my life. Knowledge is indeed power, in Bacon's proverb.

The problem may be briefly stated: What does "median mortality of eight months" signify in our vernacular? I suspect that most people, without training in statistics, would read such a statement as "I will probably be dead in eight months"—the very conclusion that must be avoided, since it isn't so, and since attitude matters so much.

I was not, of course, overjoyed, but I didn't read the statement in this vernacular way either. My technical training enjoined a different perspective on "eight months median mortality." The point is a subtle one, but profound—for it embodies the distinctive way of thinking in my own field of evolutionary biology and natural history.

We still carry the historical baggage of a Platonic heritage that seeks sharp essences and definite boundaries. (Thus we hope to find an unambiguous "beginning of life" or "definition of death," although nature often comes to us as irreducible continua.) This Platonic heritage, with its emphasis on clear distinctions and separated immutable entities, leads us to view statistical measures of central tendency wrongly, indeed opposite to the appropriateinterpretation in our actual world of variation, shadings, and continua. In short, we view means and medians as the hard "realities," and the variation that permits their calculation as a set of transient and imperfect measurements of this hidden essence. If the median is the reality and variation around the median just a device for its calculation, the "I will probably be dead in eight months" may pass as a reasonable interpretation.

But all evolutionary biologists know that variation itself is nature's only irreducible essence. Variation is the hard reality, not a set of imperfect measures for a central tendency. Means and medians are the abstractions. Therefore, I looked at the mesothelioma statistics quite differently—and not only because I am an optimist who tends to see the doughnut instead of the hole, but primarily because I know that variation itself is the reality. I had to place myself amidst the variation.

When I learned about the eight-month median, my first intellectual reaction was: fine, half the people will live longer; now what are my chances of being in that half. I read for a furious and nervous hour and concluded, with relief: damned good. I possessed every one of the characteristics conferring a probability of longer life: I was young; my disease had been recognized in a relatively early stage; I would receive the nation's best medical treatment; I had the world to live for; I knew how to read the data properly and not despair.

Another technical point then added even more solace. I immediately recognized that the distribution of variation about the eight-month median would almost surely be what statisticians call "right skewed." (In a symmetrical distribution, the profile of variation to the left of the central tendency is a mirror image of variation to the right. In skewed distributions, variation to one side of the central tendency is more stretched out—left skewed if extended to the left, right skewed if stretched out to the right.) The distribution of variation had to be right skewed, I reasoned. After all, the left of the distribution contains an irrevocable lower boundary of zero (since mesothelioma can only be identified at death or before). Thus there isn't much room for the distribution's lower (or left) half—it must be scrunched up between zero and eight months. But the upper (or right) half can extend out for years and years, even if nobody ultimately survives. The distribution must be right skewed, and I needed to know how long the extended tail ran—for I had already concluded that my favorable profile made me a good candidate for that part of the curve.

The distribution was, indeed, strongly right skewed, with a long tail (however small) that extended for several years above the eight month median. I saw no reason why I shouldn't be in that small tail, and I breathed a very long sigh of relief. My technical knowledge had helped. I had read the graph correctly. I had asked the right question and found the answers. I had obtained, in all probability, that most precious of all possible gifts in the circumstances—substantial time. I didn't have to stop and immediately follow Isaiah's injunction to Hezekiah—set thine house in order: for thou shalt die, and not live. I would have time to think, to plan, and to fight.

One final point about statistical distributions. They apply only to a prescribed set of circumstances—in this case to survival with mesothelioma under conventional modes of treatment. If circumstances change, the distribution may alter. I was placed on an experimental protocol of treatment and, if fortune holds, will be in the first cohort of a new distribution with high median and a right tail extending to death by natural causes at advanced old age.

It has become, in my view, a bit too trendy to regard the acceptance of death as something tantamount to intrinsic dignity. Of course I agree with the preacher of Ecclesiastes that there is a time to love and a time to die—and when my skein runs out I hope to face the end calmly and in my own way. For most situations, however, I prefer the more martial view that death is the ultimate enemy—and I find nothing reproachable in those who rage mightily against the dying of the light.

The swords of battle are numerous, and none more effective than humor. My death was announced at a meeting of my colleagues in Scotland, and I almost experienced the delicious pleasure of reading my obituary penned by one of my best friends (the so-and-so got suspicious and checked; he too is a statistician, and didn't expect to find me so far out on the left tail). Still, the incident provided my first good laugh after the diagnosis. Just think, I almost got to repeat Mark Twain's most famous line of all: the reports of my death are greatly exaggerated.

Before the Beginning:

Who, What, Where,
Why, When, and How?

Before I analyze data. Before I try to explain anything. Before I compute a single average or look at a single fact: *Who, What, Where, Why, When,* and *How?* Which means, establish the context. Before you get involved with the detail, ask questions: *Who* collected the data? *What* are the data about? *Where*, if that is important. *Why* were they collected? *When*, if that is important. *How* were they collected? You don't need to use a check list — ask questions.

So, for example, in a later chapter I am going to use U.S. Census data reporting the populations of states of the United States. There's the who: The U. S. Census Bureau. They have a good reputation for accuracy on total population, which is what's in these data. For some kinds of data, the results have known biases — but for these population counts, this is the best I can get. And there's the what: The data describe the population of the states of the United States. Where? The individual states. Why? To determine representation in the U. S. Congress. When? These data were published in 1991, referring to the populations in 1990. How? The census attempts to count everyone, every last person in the United States, which, strangely enough, makes the Census less accurate (not more accurate) than it would be if it used a carefully selected sample of the population.[1] I don't need the whole Who, What, Where, ... The point is to be alert and ask questions.

As a mnemonic, think of this as *step 0.* In data analysis step two is *two* variables (the relation between two variables). Step one is *one* variable — extracting information from a single variable like population size or growth rate. This is step zero, "*no* variables", the step before the analysis. Step zero is to ask whether the data is worthy of my time, whether it is trustworthy, whether it is pertinent: Who, What, Where, Why, When, and How?

[1] Curiously, a carefully drawn sample of a population can give more accurate results than an attempt to look at the entire population. The reason is a matter of cost and realism. Really, it costs a lot of money to track down every last person. So, if I talk to only one person in one hundred, I can spend one hundred times more money tracking that person down, making sure that that person is "representative" and making sure of my results for that one person. So the data from a sample can be more carefully examined at the same, or lower cost, than data from a complete enumeration. See ______in Tanur, 1989.

OPINION: Which Environmental Problems do We Think are Most Serious

	Extremely Serious	Very Serious
Hazardous and toxic waste	47%	42%
Oil spills	48	36
Air pollution	36	44
Damage to the earth’s atmosphere	39	40
Solid waste disposal	38	41
Nuclear waste	43	35
Contaminated drinking water	38	39
Destruction of forests	39	37
Threats to endangered species	26	41
Use of pesticides	22	38
World population growth	25	32
Global warming	22	34
Inefficient energy use	17	39
Reliance of fuels like coal and oil	29	34
Economic development of natural wetlands	17	33
Radon gas	11	24
Indoor air pollution	7	20

From *The Environmental Almanac*, Simon and Schuster, New York, 1992, page 11.

Figure 1

U.S Attitudes Toward Environmental Problems

Why do I ask questions? Because I'm skeptical. Because I'm careful. Why so careful? Because this is where you learn that homilies like, "don't believe everything you read" are all too valuable. To make the point, let me show you some data that failed step zero. This is data I chose *not* to analyze — let me show you why not: Preparing myself to write, I said to myself, "What would people be interested in? What am *I* interested in? Ah, let's get some data on the environment."

So I went to my local bookstore and looked around, thinking "Get some data sources that everyone can get their hands on." I looked through the almanacs, people who teach data analysis tend to collect almanacs, and there was a new one: *The 1992 Information Please Environmental Almanac*, compiled by World Resources Institute, Houghton Mifflin, 1992. Ah, I thought, just the ticket, and I thumbed through it looking for numbers.

Here's one set of numbers, reproduced in Figure 1. This is the kind of thing I was looking for. But then I remembered: "Do as you teach. You're trying to teach them that data analysis is not about numbers, it *uses* numbers. So ask questions. Where does this stuff come from? … Who, What, Where, Why, When, and How?"

"Do as you teach…", that slowed me down. Let's see, the *Almanac* tells me: "Source: Environmental Opinion Study". I wonder what that is.

Looking through the text for an answer to my question, I find it is "A 1991 poll conducted for Environmental Opinion Study, a nonprofit organization established to provide data on public attitudes on the environment…" And now I'm in trouble. Someone is trying to get past me with buzz words and puffery. The text flashes the phrase "non-profit," implying something or other. It uses the word "data", and it specifies "public attitudes". So far, the text has used a string of words to tell me the source, but the words have told me nothing.

So now I'm asking questions and I'm on full alert: When there is one loose thread in the credibility of a source, look for others. And so, looking more carefully at these data, the thing begins to unravel: Do they give me enough information so that I can find the original source and check for myself? No. Any secondary report (a report using information from another source) must give me enough information so that I can check the primary report for myself, if I choose to — but this report offers barely a clue. And now that I've seen the *Almanac* try to get past me with evasive terms, like "public attitudes", I'm even more alert. So I ask "Which public?", "Who are these people?" No answer.

More alert, I look at the numbers. Oops, the numbers are percentages. Percentages of what? … percentages of 100 people around the office of Environmental Opinion Study, percentages of a representative sample of 1,000 adults randomly sampled from the U.S. population? Percentage of what? Who knows? And I look again, noting details, noting that the vocabulary is odd. These are not the words and rhythm of standard American speech — too formal. So I wonder, how were the questions put? Did the interviewer ask "What problems do you think are serious?" Or did the interviewer ask "Do you think hazardous waste is serious?" It makes a difference: If it was the latter, then the interviewer might as well have asked whether hazardous waste is hazardous. Who could say "No" to that?

And now, as I've kept testing the credibility of these numbers, the whole thing has come apart as I look at the first row of numbers and wonder about 47% plus 42%, that is, *89%* saying toxic waste is serious? Really? *Eight-nine percent*, eighty nine out of one hundred people … of what population? Do I believe that — for any population? Frankly, no. And can I quibble with these published data? You bet I can, particularly because the writers have made it all but impossible for me to re-assure myself. So, in truth, these numbers aren't data, they're some sort of numerical decoration — taking up space. The stuff looks like data but, really, we've been asked to take the numbers on faith. And that's not the way to deal with controversial issues.

So, what's the moral of the story? Before the beginning, *Who, What, Where, Why, When,* and *How*. You do that to avoid being fooled. And when you write you must provide that information if you yourself want to be taken seriously. For all I know this “environmental opinion study” is great stuff. Maybe, somewhere in the book, there is even a footnote that answers all my questions. But, if it is great stuff, then it’s also a great pity because the authors have sabotaged their own hard work. They didn’t precede their data analysis with a solid foundation, before the beginning and, so, they might as well not have bothered with the rest.

Reading:

How to Lie with Statistics, Darrel Huff, Chapter 1, “The Sample with the Built-in Bias

________________ in Tanur, ~~ Why samples are more accurate than counts.

Description:
The Picture Worth The Thousand Words

Now, data analysis: The data have passed the test of step zero — the data are worth a look. We are about to look at the data for a single variable and to look inside the data for patterns. The key to the process is to use the best pattern recognizing device at out command: the human eyeball, with brain attached.

For this reason a large part of data analysis consists of preparing data in such a way that our native equipment can do its work. To feed intuition, we prepare graphs, the pictures that are worth a thousand words. Then we look at them and think.

Rivet your attention on this business of making pictures and avoid certain mistakes. It is a mistake to think of a picture, which is easy to look at, as less sophisticated than mathematics. That may or may not be true, depending on context. And in the context of data analysis we are looking, literally *looking* for patterns. In the trade, numbers — means, medians, measures of variation, and so forth — are referred to as “summary statistics”. And this is what they summarize, the picture, or more precisely, the pattern of the data which is made visible by a well made visual representation of the data.

You have to be careful: The eyeball may, at times, be “too good,” causing us to see patterns when they aren’t there. And the eyeball can, at times, “see” what we expect it to see, instead of what’s there. But the fact remains that the eye, the brain, and human intuition are the best tools we have for finding patterns. We label graphs because our intuition will be fed by the labels. To protect ourselves from error, we prepare we charts and graphs from which errors will stand out visually, as breaks from a pattern — and be corrected.

So we begin by looking at the data, beginning with one variable, and using a technique known as the “Stem and Leaf”.[2]

[2] From John Tukey’s *Exploratory Data Analysis*, Chapter 1, “Scratching Down Numbers”, Addison Wesley, 1977.

From *FOOD VALUES OF PORTIONS COMMONLY USED*, BOWES & CHURCH, 1975, p. 11

4.1 CEREALS (A) - READ TO SERVE

Food	WT	CAL	CHO	FAT	TRP	LEU	LYS	MET	Na	Ca	P	THI	NIA	VIA
Measure		PRO	FIB	FAP	PHA	ISL	VAL	THR	K	Mg	Fe	RIB	ASC	VID
	gm	gm	gm	gm	mg	mg	mg	mg	mg	mg	mg	mcg	mg	iu
BARLEY CEREAL, GERBER'S	36	128	27.3	.2	54	299	145	62	215	231	260	1015	5.1	(0)
1 cup		4.3	.4		222	183	216	145	149		18	763	0	
BRAN, ALL-, KELLOGG'S	28	95	21.4	.7					370	24	350	110	5.0	(0)
1/2 cup		3.1	2.3								2.9	90	0	400
BRAN FLAKES, 40%, KELLOGG'S	28	101	22.6	.6					340	16	170	100	2.4	(0)
3/4 cup		2.9	1.0								1.3	50	0	
BRAN FLAKES, 40%, POST'S	28	100	22.0	.5					(340)	-	110	130	1.5	(0)
3/4 cup		2.8	1.0								1.0	-	-	0
BRAN, RAISIN, KELLOGG'S	21	73	16.6	.4					280	13	105	63	1.5	
1/2 cup		1.8	-								1.0	-	0	0
BRAN, RAISIN, POST'S	28	99	22.0	.4							94	100	1.1	
2/3 cup		2.2	-								1.0	-	0	0
CHEERIOS, GENERAL MILLS'	25	102	17.7	1.8					275	42	100	302	.5	
1 cup		3.4	.3								1.1	49	0	0
CORN FETTI, POST'S	28	110	25.0	.1						-	-	110	.5	
3/4 cup		1.5									.4	-	0	0
CORN FLAKES (b)	25	95	21.0	.1	14	272	40	35	165	6	16	100	.5	(0)
1 cup		2.1	.2		92	80	100	72	40		.5	20		0
CORN SOYA SHREDS	28	103	21.0	.1					310	24	52	190	.6	(0)
3/4 cup		5.1	.3								1.2	40	0	0
GRAPE NUTS (c)	28	110	24.0	.2	-	196	45	39	17	-	-	130	1.5	(0)
1/4 cup		2.8			137	137	137	90			1.0	-	0	0
GRAPE NUT FLAKES	28	110	23.0	.4						-	-	130	1.6	(0)
3/4 cup		2.7									1.2	-	-	-
HIGH PROTEIN CEREAL, GERBER'S	29	102	14.5	.3					226	213	241	818	4.1	
3/4 cup		10.2	.4						313		14.	615		
HI PRO, GENERAL MILLS'	21	80	14.1	.3					294	65	84	345	3.1	
1cup		4.8	.1								3.9	432		
KIX, GENERAL MILLS'	25	99	20.2	1.0					275	5	22	214	.7	(0)
1 cup		2.0	.1								1.5	46		0
KRUMBLES, KELLOGG'S	28	103	23.8	.3					170	11	110	10	2.0	(0)
3/4 cup		2.6									1.0	30	0	0
MIXED CEREAL, GERBER'S	21	76	15.4	.3					126	111	133	592	3.0	
1/2 cup		3.0	.2						72		10.5	445		
MUFFETS, QUAKER	23	80	18.2	.3					1.0	10	87	50	1.0	(0)
1 biscuit		2.2	.6								.9	20	0	0
OATMEAL (d)	27	98	18.6	.6	58	337	165	66	135	153	169	761	3.8	(0)
3/4 cup		4.5	.4		240	232	267	149	101	13.5	572			0
POST TOASTIES	28	100	24.0	.1						-	-	110	.5	(0)
1 1/4 cup		2.1	.2								.4	-	-	0
RICE CEREAL, GERBER'S	27	97	21.9	.4					205	179	171	761	3.8	
3/4 cup		1.5	.1						56		13.5	572		
RICE FLAKES	32	123	27.4	.1	16	-	20	-	311	11	53	110	1.4	(0)
1 cup		2.1	.3		102	-	-	-			.6	20		0
RICE KRISPIES, KELLOGG'S	28	107	25.1	.1					280	7	33	110	2.0	(0)
1 cup		1.6									.5	10	0	0
RICE, PUFFED, QUAKER	13	51	11.5	.1					.3	2	13	60	.6	(0)
1 cup		.8	.1								.2	10	0	0
SPECIAL K CEREAL, KELLOGGS'	16	60	12.5	.1					193	17	41	228	2.9	
1 cup		3.2									2.5	285	6	228
WHEAT FLAKES, QUAKER	36	125	28.0	.4	49	363	147	52	403	17	108	160	1.9	(0)
1 cup		4.4	.5		195	202	233	145			1.3	60	0	0
WHEATIES, GENERAL MILLS'	28	104	22.5	.6					392	11	78	167	1.6	(0)
1 cup		2.8	.5								1.7	47	0	0
WHEAT, PUFFED, QUAKER	12	43	9.5	.2					1	3	40	70	.9	(0)
1 cup		1.6	.2								.5	30	0	0
WHEAT, SHREDDED	22	84	18.3	.3	18	149	72	30	.5	11	93	65	1.0	(0)
1 biscuit		2.2	.5		105	98	126	88			.8	23	0	0
WHEAT CHEX, RALSTON	28	102	23.4	.3					225	11	105	40	1.5	
1/2 cup, 47 biscuits		2.8	.6								.9	60		

(a) All the cereals listed on this page may be served from the package without further prepartion. When served with milk or cream and/or sugar the addenda should be consulted.

(b) The amino acid values are from reference 15. Sodium and potassium figures are calculated from reference 14 aned 14a.

(c) These values for amino acids are derived from reference 15.

(d) The ready-to-serve product is indicated here. Amino acid data is from refeence 15.

NOTE: A serving of cereal varies with individual taste, age, and activity level. A common size serving is 1 ounce.

Stem and Leaf

Technique

What's for breakfast? I'm in one of my fitness moods, it's 5 A.M. and the first question of the day is "What's for breakfast?" I want a high protein breakfast. And, more generally, I want to know what it is that makes one breakfast cereal different from another. My data are from *Food Values of Portions Commonly Used*, by Bowes and Church, a dietitian's handbook whose introduction is chock full of references that tell me where the data come from should I want to check for myself.[1] Here is their table of data on cereals, ready to serve, describing nutrients found in a standard portion, Figure .1

Reproduce page from Bowes and Church, page 11, on facing page.

Reading the top pair of lines, the table indicates that Gerber's barley cereal is usually served in a one cup portion weighing thirty-six grams. It provides 128 calories, 4.3 grams of protein, 27.3 grams of carbohydrates, .4 grams of fiber, .2 grams of fat, and negligible grams of polyunsaturated fat (FAP). Reading across, the data indicate the amounts of eight amino acids, in milligrams, of six minerals, in milligrams, and of six vitamins, in various units.

Focus on the grams of protein: I want to know how much protein these breakfast cereals tend to provide, what's high, what's low, and more generally what it is about "breakfast" cereal that leads some to

[1] The frequently updated edition is by Pennington and Church, published by Harper & Row. I'm using an out-of-date, 1975, edition, page 11, because it was less complete than more-current editions and therefore easier to use as an exercise. The up-to-date edition includes more information on more products.

be high in protein while others are low? What is the "mechanism" behind the facts?

O.K., that's the agenda, but first some technique — the real agenda for this chapter. And the first technique, not my breakfast, is called the "Stem and Leaf."

This is a technique that is both extremely useful and extremely modest — hard to take seriously until it "works" for you, time after time — too simple to pay off, but it does. It is also one of those techniques that falls on the unseen side of the data analysis — rarely seen in a final report, but often found on the scratch pad of the data analyst, usually in pencil, usually with notes and scratchings all over it: It is informal and extremely useful.

Let me begin simply and mechanically, without context, by simply extracting the protein numbers from Figure 1 and illustrating the technique.

4.3, 3.1, 2.9, 2.8, 1.8, 2.2, 3.4, 1.5, 2.1, 5.1, 2.8, 2.7, 10.2, 4.8, 2.0, 2.6, 3.0, 2.2, 4.5, 2.1, 1.5, 2.1, 1.6, .8, 3.2, 4.4, 2.8, 1.6, 2.2, 2.8

Figure 2
Numbers, for Practice, Extracted from Protein Values of Figure 1.

Mechanically, beginning with these numbers, a Stem and leaf is a new copy of the same numbers — but re-grouped by size and presented in a way that shows the "shape" of the data. It's a first step in engaging your intuition and experience as allies in the process of data analysis.

Working just with the numbers, Figure 2 is a stem and leaf for the numbers in Figure 1.

0.8

0	8	*← Stem for numbers greater than or equal to 0 but less than 1*
1	85566	*← Stem for numbers greater than or equal to 1 but less than 2*
2	98218706211828	
3	1402	
4	3854	
5	1	
10	2	

Figure 2
Stem and Leaf, Integer Stems

The "stems" identify ranges, dividing the numbers into major divisions. The "leaves" identify each of the numbers within the division: Leaves are the numbers attached to the stems. Here, in this first example, stems divide the data using digits to the left of the decimal, with stems 0 through 10. The leaves mark each datum on the stem with an additional digit. So, the number 0.8 is represented in stem 0, identified by the leaf "8". And the number 10.2 is represented in stem "10", identified by the leaf "2". All of the numbers greater than or equal to 0, but less than 1, are represented in the first stem. All of the numbers greater than or equal to 1, but less than 2, are represented in the second stem. The leading digits, "0", "1", "2", and up to "10" label the stems. The final digits identify the leaves.

Altogether, the effect of the stem and leaf, when it is completed, is to put the data in rank order, roughly, and to show the shape of the distribution of the numbers. Much of the value of the stem and leaf lies in the process as much as the result: In the process you almost literally *feel* which numbers are typical as you record the leaves one by one, putting them in place. By the time you are done you "know" your data — which numbers are small, which numbers are large compared to the rest, which ones stand out from the others flagging that, for these few numbers something is different — they don't belong with the rest. In the process, you begin to get a feel for the numbers.

When you are done you have a picture. And what you do with the picture is stand back and look at it for a moment: Here, for these thirty numbers, the shape is symmetrical with many numbers in the middle, and few at either end. Separately, one of the thirty numbers is out on its own. That symmetrical shape, many numbers in the middle and a few at either end, is what's known as a bell-shaped distribution — some data show it, some data don't. And the one suspicious case at the end is what's known as an outlier.

That's a stem and leaf for these numbers. But usually it's not quite that simple. Usually you have to try a couple of different stem and leaf drawings before you get one that looks right. What does it mean to "get one that looks right?" To show you what I mean, let me practice with these numbers.

For practice, I can use different stems that expand the stem and leaf as in Figure 3. Here I've expanded the whole thing twice as far physically: I've used 0.0 through 0.4 for the first stem, 0.5 through 0.9 for the second stem, and so forth, giving me twice as many stems.

0.8

0	
0.5	8
1	
1.5	85566
2	2102112
2.5	9887688
3	1402
3.5	
4	34
4.5	85
5	
10	2

Figure 3
Stem and Leaf, Expanded

Or, I can compress the stem and leaf by dividing at 0, 2, 4, 6, 8 and 10, as in Figure 4.

0.8 1.8

0	8:85566
2	98218706211828:1402
4	3854:1
6	
8	
10	2

Figure 4
Stem and Leaf, Compressed

If you expand the data too much, then the shape gets irregular—showing gaps as in Figure 3. Alternatively I can try a third set of stems that compress the data. And if you compress the data too much then the shape disappears into a lump. How many stems should you use? Try for 5 to 10 but there is no fixed answer. Look for a shape that is fairly compact, like Figure 2. If you expand it and the shape begins to break up, as in Figure 3, then you've gone to far. Here, Figure 2 is good enough.

One not-so-minor detail of the stem and leaf that should be made explicit is the labeling. Labels are as much a part of the technique as the numbers. And the rule for labeling is: Make it Clear. Labels need to identify the stems, they need to identify the leaves, and— when we get to real data—they have to describe the unit of analysis (e.g., kilograms of protein or grams of protein or milligrams of protein). Briefly, the labels are an echo of the "Who, What, Where, Why, When, and How" The specify what the data are about.

For the stems there are at least three styles I could have used to label the expanded stems in Figure 3. In Figure 4 I used digits. Alternatively, as in Figure 5, I could have simply repeated the leading digit, clear enough in context, or repeated the leading digit and marked the second case with a "*", to distinguish it from the other.

0	0
0	0*
1	1
1	1*
2	2
2	2*
3	3
3	3*
4	4
4	4*
5	5
5	5*

Figure 5
Stem and Leaf, Alternative Labels for Stems

Which style should you use? Actually I myself used all three of these styles as I worked with these data. First I used the labels on the left of Figure 5 — they're the easiest — using each label twice. But then, as I worked with this style, I didn't like it: For some reason, I found these stems easy to write down but hard to use: Using these stems I kept making errors by attaching the leaves to the wrong stems. So I changed to the “*” version in order to differentiate the stems. But then, as I used it, I kept making errors — still putting the leaves on the wrong stems. So I changed again, settling on the labels 0, 0.5, … of Figure 2. With the limits of the stem visible on the page, built-in to the labels for the stems, I was able to work faster and with fewer errors.

And I give you this blow by blow summary of my thinking — first I did this and then I did that — just to make it clear how you decide, and how I decide, to do the stem and leaf: Don't look for the “one true way” of doing it. There is none. Instead, think of what you are trying to accomplish and feel free to change technique until it works.

Also note that I put a label on each of the completed stem and leaf diagrams. a “0.8” with an arrow attached. That was for my benefit, to help me read the stem and leaf when I have to go back to make sense of my own work — an hour, or a day or a month later. And you should use it yourself so that your own stem and leaf drawings can be read: Labels are *not* an after-thought. They are part of good technique.

Exercises

1. Practice several forms of the stem and leaf on the numbers for fat content in grams

.2, .7, .6, .5, .4, .4, 1.8, .1, .1, .1, .2, .4, .3, .3, 1, .3, .3, .3, .6, .1, .4,.1, .1,.1, .1, .4, .6, .2, .3, .3

and, again, on the numbers of calories for the same breakfast cereals

128, 95, 101, 100, 73, 99, 102, 110, 95, 103, 110, 110, 102, 80, 99, 103, 76, 80, 98, 100, 97, 123, 107, 51, 60, 125, 104, 43, 84, 102

2. Use a pair of dice and construct the stem and leaf drawing for one hundred passes with the dice. Before you begin, ask questions: What do expect the diagram to look like? Why ? Now, throw the dice. What do you get and why?

Application: Protein Content of Breakfast Cereals

Now, what's for breakfast? For *data* analysis — with emphasis on the word *data* — I have to place these numbers back in context: The numbers record the grams of protein in "commonly used portions" of breakfast cereals. Now, in context, there is a difference: Here is where

I get to assert the advantages of my human brain as compared to the mechanical "brain" of a computer. In context this human brain has expectations, and a little experience with breakfast, and I intend to use those advantages in the course of my analysis. I, the analyst with expectations, expect the quality of the food to depend on the ingredients and also upon the manufacturer and the process. In context I expect these cereals to divide into two batches, good stuff and bad stuff, or, health foods and junk foods.

Having thought about what I expect, I'm ready to begin my stem and leaf. What kind of stems? Well, as before, when I looked at the numbers, I see numbers like 4.3, 3.1, and 2.9, reading down. So, again, I'll start with stems indicating grams of protein in one gram intervals.

```
0|
1|
2|
3|
4|
5|
6|
 |
```

Protein in Grams

O.K., now what about the leaves? You've seen these numbers before, as numbers. Now, in context, these are data: They have been identified with something real and that makes a difference. Bearing in mind that I'm using the numbers to get at something else, something about breakfast cereals, I'm going to use leaves that advance my purpose: I'm *expecting* the ingredients to tell part of the story so, somehow, the ingredients should be marked, in the leaves. And again, I'm *expecting* that some manufacturers make a better product. So, I want to keep track of manufacturers. And then, for bookkeeping purposes, I note that the data come in alphabetical order. That's useful: I want to use that alphabetical order in order to be able to connect my summary of the data, in the stem and leaf, back to the full data that it comes from. All together, there's a lot of information here, waiting to be organized. So, I'm going to use labels for the leaves, not numbers. And I'm going to build into those labels whatever useful information I can manage (in addition to the grams of protein). So I begin

0	
1	
2	Bran Fl Kell;
3	Bran Kell;
4	Barley Gerb;
5	
6	

Stem and Leaf
First Three Leaves

Protein in Grams

Those are the first three leaves and, pausing for the moment, let me note a few things: For one, the physical lengths of these three stems are not quite equal, so the visual shape of the stem and leaf will be a little distorted. True, but that's not too important. For another, note that I changed my stem labels a little, on the fly. The second cereal was "Bran, All-, Kellogg's," in the data, and I wrote down "Bran Kell," in the leaf. But then the third cereal entry was also "Bran Kell", like the second, except that this third food is a flake and the second was not. Ah, that's new information, at no extra cost — something about the process. So I'll put that information into the third leaf, even though I didn't use it in the second leaf: Consistency is nice, but not when it gets in the way. (And if this becomes important, the textual stems, unlike numerical stems, will make it easy to go back to the data for more detail.) Continuing

0	
1	Bran Rais Kell;
2	Bran Fl Kell; Bran Fl Post;Bran Rais Post;
3	Bran Kell; Cheerios GM
4	Barley Gerb;
5	
6	

Stem and Leaf
First Six Leaves

Protein in Grams

Those are the first seven leaves and I've got trouble again: the data themselves are not always labeled the same way and my seventh datum, for "Cheerios," doesn't specify ingredients. What do I do? I use it anyway. I use what I've got. And if the missing information becomes important, I can find out later. Continuing:

0	
1	Bran Rais Kell; Corn F Post;
2	Bran Fl Kell; Bran Fl Post; Bran Rais Post; Corn Fl; Grape Nuts; Grape Nut Fl
3	Bran Kell; Cheerios GM
4	Barley Gerb;
5	Corn Soya;
6	
7	
8	
9	
10	High Pro Gerb;

Stem and Leaf
First Eleven Leaves

Protein in Grams for Commonly Used Portions of Breakfast Cereal
Example: High Protein Gerbers 10.2 grams

Thirteen items into the procedure, and now there's a big one, out of line with the rest: High Pro Gerbers has several times more protein than the competition. I could stop now and think about it, but there's not much more data, so I'll continue.

0	Rice Puff Q;
1	Bran Rais Kell; Corn F Post; Rice Gerb; Rice Kr Kell; Wheat Puff Q
2	Bran Fl Kell; Bran Fl Post; Bran Rais Post; Corn Fl; Grape Nuts; Grape Nut Fl; Kix GM; Krumb Kell; Muff Quak; Post T; Rice Fl; Wheats GM; Whet Shred; Wheat Chex Ralst
3	Bran Kell; Cheerios GM; Mixed Gerb; Spec K Kell
4	Barley Gerb; High Pro GM; Oatml; Wheat Fl Quak
5	Corn Soya;
6	
7	
8	
9	
10	High Pro Gerb;

Stem and Leaf
Breakfast Cereals

Protein in Grams for Commonly Used Portions of Breakfast Cereal
Example: High Protein Gerbers 10.2 grams

O.K., that's it for the moment: I ran out of room and ruined the shape on stem "2". I've still got one real stand out, "High Pro Gerb". And I've been a little inconsistent, paying more attention to the process: flaked, puffed, or shredded, than I had intended. This is probably good enough. But I probably could have figured out pretty early that I should have been using different stems: I used eleven stems because eleven gave me a convenient division of the range between zero and ten. But with these stems things are bunching up in part of the range, while almost half of the stems, between five and ten are nearly empty. I might have been better off expanding the stem and leaf within the range from zero to five, just leaving the one very high protein cereal, Gerbers High Pro, out there on its own.

0	
0.5	Rice Puff Q;
1	
1.5	Bran Rais Kell; Corn F Post; Rice Gerb; Rice Kr Kell; Wheat Puff Q
2	Bran Rais Post; Corn Fl; Kix GM; Muff Quak; Post T; Rice Fl; Whet Shred;
2.5	Bran Fl Kell; Bran Fl Post; Grape Nuts; Grape Nut Fl;Krumb Kell; Wheats GM; Wheat Chex ; Ralst
3	Bran Kell; Cheerios GM; Mixed Gerb; Spec K Kell
3.5	
4	Barley Gerb; Wheat Fl Quak
4.5	High Pro GM; Oatml;
5	Corn Soya;
4.5	
10	High Pro Gerb;

Stem and Leaf Expanded

Protein in Grams for Commonly Used Portions of Breakfast Cereal
Example: High Protein Gerbers 10.2 grams

Trying out the expanded stem and leaf, there's not much difference. Either way, by grams or expanded to half-gram intervals, it's clear that the most frequent values are in the neighborhood of 2.5 grams of protein. And, still, the prominent event is the outlier, High Pro Gerber's, with protein that is four times the typical value.

Now, stepping back to look at these things, looking at either one of the stem and leaf diagrams, what can I learn about cereal? For one thing, I learn that one of my expectation was wrong: I expected something simple, good food versus bad food. Wrong: It's not that simple. With the exception of Gerber's, most of the foods form one nice batch, one nice distribution not two batches, good and bad, but one. And while "bran cereals" get lots of good public relations as "health foods", they are not notably high in protein. "Good" versus "bad" is too simple.

However, while the twenty-nine cereals (other than Gerbers) are not divided into two types, good versus bad, the range of values shown in the stem and leaf diagram shows that there *is* a very large variation within this group. How large? The protein values range from one cereal in the 0-gram stem to one cereal in the 5-gram stem. Precisely, how large? Using the alphabetical cues in the stem and leaf, I can quickly fill-in this detail by going back to the data for more information: The stem and leaf diagram displays extremes “Rice Puff Q” and “Corn Soya”, which allows me to glance back to the data for the full numbers, which are 0.8 and 5.1. So, how large is the variation? With the exception of one outlier, the protein values range from 0.8 to 5.1 grams of protein per serving, with the high protein cereals (at the end of the range) providing six times the protein content of the lowest value — a big contrast.

Now, I wanted to know “Why?” Why do some cereals have high protein. What about High Pro Gerbers? If I’m looking for high protein this is the first place to begin. It has unusually high protein, extremely high. Why? Perhaps it is the manufacturer. Is there something about Gerber that I should favor, as a brand? That’s one hypothesis — it’s the manufacturer. And I can check that hypothesis by going back through the leaves, marking the manufacturer’s names: (Here I’ve marked off Gerber with bolding. Ordinarily, I’d circle the Gerber’s in my existing stem and leaf, or mark them with a bright color.) Looking at the marked-up stem and leaf, I see Gerber all over, low, medium, and high. So, it’s not that simple. Cross-off that hypothesis.

0	
0.5	Rice Puff Q;
1	
1.5	Bran Rais Kell; Corn F Post; **Rice Gerb**; Rice Kr Kell; Wheat Puff Q
2	Bran Rais Post; Corn Fl; Kix GM; Muff Quak; Post T; Rice Fl; Whet Shred;
2.5	Bran Fl Kell; Bran Fl Post; Grape Nuts; Grape Nut Fl; Krumb Kell; Wheats GM; Wheat Chex Ralst
3	Bran Kell; Cheerios GM; **Mixed Gerb**; Spec K Kell
3.5	
4	Barley Gerb; Wheat Fl Quak
4.5	High Pro GM; Oatml;
5	Corn Soya;
4.5	
10	**High Pro Gerb;**

Stem and Leaf
Highlighting Gerber

Protein in Grams for Commonly Used Portions of Breakfast Cereal
Example: High Protein Gerbers 10.2 grams

Let me try again. Maybe it is Gerber, but complicated by the choice of ingredients:. Maybe Gerber's Rice is higher protein than other people's rice. Highlighting again, to check my hunch: No, Gerber's rice is right in there among the other rice cereals. But note, before I go on, how easy it was to do these checks of my hunches, or "hypotheses", and how hard it would have been if I had just recorded the digits (of Figure __). And note the rudiments of scientific reason built in to these last few steps: I used my expectations to form testable hypotheses. I formulated the display in order to test the hypotheses. I tested them. And, so far at least, both appear to be false.

0	
0.5	**Rice Puff Q**;
1	
1.5	Bran Rais Kell; Corn F Post; **Rice Gerb**; **Rice Kr Kell**; Wheat Puff Q
2	Bran Rais Post; Corn Fl; Kix GM; Muff Quak; Post T; **Rice Fl**; Whet Shred;
2.5	Bran Fl Kell; Bran Fl Post; Grape Nuts; Grape Nut Fl; Krumb Kell; Wheats GM; Wheat Chex Ralst
3	Bran Kell; Cheerios GM; Mixed Gerb; Spec K Kell
3.5	
4	Barley Gerb; Wheat Fl Quak
4.5	High Pro GM; Oatml;
5	Corn Soya;
4.5	
10	High Pro Gerb;

Stem and Leaf
Highlighting Rice

Protein in Grams for Commonly Used Portions of Breakfast Cereal
Example: High Protein Gerbers 10.2 grams

Now I've used up everything I know about Gerbers High Pro, the extreme case, trying to figure out the "mechanism" that makes it special. No luck. So, having used up what I know about Gerbers, I'll have to look elsewhere. I'll go to the next largest value and see what this one might tell me: It says "Corn Soya", labeled by ingredients. Does this tell me anything? Comparing this second highest protein cereal to the lowest protein cereal my leaves show me "Corn Soya" versus "Rice Puff". That suggests a possibility, perhaps the important feature is the combination of ingredients, corn and soy, at one extreme versus rice at the other? To check that out I'll go back to the stem and leaf, making it up again. (Ordinarily I'd go back through the same stem and leaf with more colored pens or mark it with some fancy symbol, marking each ingredient where it is known. Here, I'll use different type fonts.)

0	
0.5	**Rice** Puff Q;
1	
1.5	*Bran* Rais Kell; Corn F Post; **Rice** Gerb; **Rice** Kr Kell; Wheat Puff Q
2	*Bran* Rais Post; Corn Fl; Kix GM; Muff Quak; Post T; **Rice** Fl; Wheat Shred;
2.5	*Bran* Fl Kell; Bran Fl Post; Grape Nuts; Grape Nut Fl; Krumb Kell; Wheats GM; Wheat Chex Ralst
3	*Bran* Kell; Cheerios GM; Mixed Gerb; Spec K Kell
3.5	
4	Barley Gerb; Wheat Fl Quak
4.5	High Pro GM; Oatml;
5	Corn Soya;
4.5	
10	High Pro Gerb;

Stem and Leaf
Highlighting Ingredient

Protein in Grams for Commonly Used Portions of Breakfast Cereal
Example: High Protein Gerbers 10.2 grams

That wasn't what I expected. Generalizing from two cases, I expected, or hoped to find that corn (underlined) is high in protein and notably separate from rice (bold). But the corn and rice are roughly in the same range — except in the one case where the corn was mixed with soy. *Other* contrasts, other than corn versus rice do look promising: There is a suggestion that the wheat cereals are higher protein than rice cereals. So I'll follow up on the "ingredients hypothesis", with a new set of stem and leaf drawings. For this hypothesis I can use a pair of stem and leaf drawings, back to back, separated by ingredient. For example, wheat versus rice:

	0	
	0.5	**Rice** Puff Q;
	1	
Wheat Puff Q	1.5	**Rice** Gerb; **Rice** Kr Kell
Whet Shred	2	**Rice** Fl;
Wheats GM; Wheat Chex Ralst	2.5	
	3	
	3.5	
Wheat Fl Quak	4	
	4.5	
	5	
	4.5	

Stem and Leaf
Wheat Versus Rice

That looks good and I'm feeling more certain of my hypothesis. I'll venture a guess now: Watch the ingredient. Guess that wheat cereals are higher protein than rice cereals. And, noting that corn cereal is pretty much like rice cereal *except for the one case in which the corn is mixed with soy* — I'll guess that soy beans are the key to the high protein content of corn soya. And that in turn leads me to a guess about Gerbers High Protein.

O.K. now, stepping aside from the data analysis, using the stem and leaf drawings, what have I got: I have a hunch that the main ingredient is the best predictor of protein content (not the manufacturer), plus a reasonable hypothesis explaining the anomalous Corn Soya (it has too much protein for a corn cereal), plus a guess that could explain Gerber's with its remarkably high protein. Have I proven anything? Have I proved, statistically, that the wheat distribution is different from the rice distribution — beyond a statistical doubt? No, nor do I need to. I've combed these data for ideas, detective style. I've eliminated some bright ideas that turned out to be poor and I've moved toward one idea that looks good, so far. That's my data analysis. And, since I'm squeezing these data to learn something about the breakfast cereals — remembering my goal — there's an easy way to check my hunch: Find Gerbers in the store and read the box. Sure enough, moving my data analysis laboratory to the nearest grocery store, the *only* ingredient in High Protein Gerbers is finely sliced soybeans. Q.E.D.

Suggested reading: “They Should have Used A Shovel”, David Freedman, *Sociological Methodology*??, __________

Exercises

1. Using the examination of protein content as a model, examine the distribution of fat content and calories for these cereals.

2. The “Dow Jones Industrial Average” is a weighted average of the prices for shares of stock for thirty companies. The data below show the change in price for each of the thirty companies. Construct a stem and leaf diagram for the change in price during the week March 18 to March 25 (column 5). Discuss. **(Check your hand outs for more recent prices)**

	Dates:	March 18, 1994	March 25, 1994	Change	Change as Per Cent of March 18th Value
	Dow Jones Average of Thirty Industrials	3,895.65	3,774.73	- 120.92	-3.10%
		Price per Share of Stock	Price per Share of Stock		
1	Alcoa Aluminum	$77.13	$76.25	– $0.88	-1.13%
2	Allied Signal	$77.00	$76.50	– $0.50	-0.65%
3	American Express	$30.13	$29.63	– $0.50	-1.66%
4	AT&T	$53.50	$52.63	– $0.88	-1.64%

5	Bethlehem Steel	$21.50	$20.88	– $0.63	-2.91%
6	Boeing	$47.00	$45.75	– $1.25	-2.66%
7	Caterpillar	$119.13	$116.63	– $2.50	-2.10%
8	Chevron	$92.75	$89.50	– $3.25	-3.50%
9	Coca Cola	$41.88	$41.75	- $0.13	-0.31%
10	Disney	$47.00	$45.00	– $2.00	-4.26%
11	duPont	$58.75	$56.25	– $2.50	-4.26%
12	Kodak	$45.13	$44.75	– $0.38	-0.83%
13	Exxon	$65.88	$65.50	– $0.38	-0.57%
14	General Electric	$104.50	$102.13	– $2.38	-2.27%
15	General Motors	$59.88	$56.88	– $3.00	-5.01%
16	Goodyear	$45.25	$41.75	– $3.50	-7.73%
17	IBM	$57.13	$54.00	– $3.13	-5.48%
18	International Paper	$70.25	$66.38	– $3.88	-5.52%
19	MacDonalds	$60.88	$58.38	– $2.50	-4.11%
20	Merck	$31.75	$30.13	– $1.63	-5.12%
21	MMM	$103.00	$100.00	– $3.00	-2.91%
22	Morgan	$64.75	$63.88	– $0.88	-1.35%
23	Philip Morris	$55.00	$51.50	– $3.50	-6.36%
24	Proctor & Gamble	$56.63	$53.75	– $2.88	-5.08%
25	Sears	$48.13	$46.00	– $2.13	-4.42%
26	Texaco	$66.25	$65.88	– $0.38	-0.57%
27	Union Carbide	$25.88	$25.00	– $0.88	-3.38%
28	United Technologies	$68.25	$66.50	– $1.75	-2.56%
29	Woolworth	$19.88	$19.13	– $0.75	-3.77%
30	Westinghouse	$13.25	$13.00	– $0.25	-1.89%

	Dates:	March 3, 1995	March 10, 1995	Change	Change as a Per Cent of the March3rd Value
	Dow Jones Average of Thirty Industrials	3989.61	4035.61	46	1.15%
		Price per Share of Stock	Price per Share of Stock		
1	Alcoa Aluminum	$78.50	$76.50	-2.00	-2.55%

2	***Allied Signal***	***$37.75***	***$38.00***	***0.25***	***0.66%***
3	***American Express***	***$33.38***	***$32.63***	***-0.75***	***-2.25%***
4	***AT&T***	***$51.25***	***$52.13***	***0.88***	***1.72%***
5	***Bethlehem Steel***	***$15.38***	***$15.13***	***-0.25***	***-1.63%***
6	***Boeing***	***$46.25***	***$46.88***	***0.63***	***1.36%***
7	***Caterpillar***	***$49.00***	***$49.88***	***0.88***	***1.80%***
8	***Chevron***	***$46.88***	***$48.00***	***1.12***	***2.39%***
9	***Coca Cola***	***$55.13***	***$56.75***	***1.62***	***2.94%***
10	***Disney***	***$53.75***	***$56.13***	***2.38***	***4.43%***
11	***duPont***	***$55.38***	***$55.38***	***0.00***	***0.00%***
12	***Kodak***	***$51.38***	***$51.88***	***0.50***	***0.97%***
13	***Exxon***	***$63.25***	***$65.00***	***1.75***	***2.77%***
14	***General Electric***	***$53.00***	***$54.75***	***1.75***	***3.30%***
15	***General Motors***	***$39.88***	***$41.63***	***1.75***	***4.39%***
16	***Goodyear***	***$37.13***	***$36.00***	***-1.13***	***-3.04%***
17	***IBM***	***$79.88***	***$81.13***	***1.25***	***1.56%***
18	***International Paper***	***$73.50***	***$73.00***	***-0.50***	***-0.68%***
19	***MacDonalds***	***$33.00***	***$33.88***	***0.88***	***2.67%***
20	***Merck***	***$41.63***	***$41.88***	***0.25***	***0.60%***
21	***MMM***	***$54.38***	***$56.13***	***1.75***	***3.22%***
22	***Morgan***	***$65.38***	***$63.25***	***-2.13***	***-3.26%***
23	***Philip Morris***	***$62.00***	***$63.38***	***1.38***	***2.23%***
24	***Proctor & Gamble***	***$66.13***	***$67.25***	***1.12***	***1.69%***
25	***Sears***	***$50.38***	***$50.75***	***0.37***	***0.73%***
26	***Texaco***	***$63.75***	***$65.13***	***1.38***	***2.16%***
27	***Union Carbide***	***$28.00***	***$27.75***	***-0.25***	***-0.89%***
28	***United Technologies***	***$66.00***	***$66.13***	***0.13***	***0.20%***
29	***Woolworth***	***$15.63***	***$15.88***	***0.25***	***1.60%***
30	***Westinghouse***	***$14.75***	***$14.63***	***-0.12***	***-0.81%***

3. Ditto for 1995, March 3 to March 10 (column 5).

4. Ditto for change during the one year, approximately, between March 18, 1994 and March 10, 1995. Before you begin: What do you expect? Why. And then when you see the Stem and Leaf: What did you find? What questions did it raise? Why? Explain — "What's going on here?"

5. Construct a stem and leaf diagram for the states of the United States with respect to their rates of infant mortality.

	Division and State	*Region*	*Total Infant Mortality Rate, 1988*
1		**U.S.**	**10.0**
2		**N.E.**	**8.1**
3	Maine	N.E.	7.9
4	New Hampshire	N.E.	8.3
5	Vermont	N.E.	6.8
6	Massachusetts	N.E.	7.9
7	Rhode Island	N.E.	8.2
8	Connecticut	N.E.	8.9
9		**M.A.**	**10.3**
10	New York	M.A.	10.8
11	New Jersey	M.A.	9.9
12	Pennsylvania	M.A.	9.9
13		**E.N.C.**	**10.5**
14	Ohio	E.N.C.	9.7
15	Indiana	E.N.C.	11.0
16	Illinois	E.N.C.	11.3
17	Michigan	E.N.C.	11.1
18	Wisconsin	E.N.C.	8.4
19		**W.N.C.**	**8.9**
20	Minnesota	W.N.C.	7.8

21	Iowa	W.N.C.	8.7
22	Missouri	W.N.C.	10.1
23	North Dakota	W.N.C.	10.5
24	South Dakota	W.N.C.	10.1
25	Nebraska	W.N.C.	9.0
26	Kansas	W.N.C.	8.0
27		**S.A.**	**11.6**
28	Delaware	S.A.	11.8
29	Maryland	S.A.	11.3
30	Dist Columbia	S.A.	23.2
31	Virginia	S.A.	10.4
32	West Virginia	S.A.	9.0
33	North Carolina	S.A.	12.5
34	South Carolina	S.A.	12.3
35	Georgia	S.A.	12.6
36	Florida	S.A.	10.6
37		**E.S.C.**	**11.4**
38	Kentucky	E.S.C.	10.7
39	Tennessee	E.S.C.	10.8
40	Alabama	E.S.C.	12.1
41	Mississippi	E.S.C.	12.3
42		**W.S.C.**	**9.4**
43	Arkansas	W.S.C.	10.7
44	Louisiana	W.S.C.	11.0
45	Oklahoma	W.S.C.	9.0
46	Texas	W.S.C.	9.0
47		**Mt.**	**9.2**
48	Montana	Mt.	8.7
49	Idaho	Mt.	8.8
50	Wyoming	Mt.	8.9
51	Colorado	Mt.	9.6
52	New Mexico	Mt.	10.0

53	Arizona	Mt.	9.7
54	Utah	Mt.	8.0
55	Nevada	Mt.	8.4
56		**Pac**	**8.6**
57	Washington	Pac	9.0
58	Oregon	Pac	8.6
59	California	Pac	8.6
60	Alaska	Pac	11.6
61	Hawaii	Pac	7.2

Total Infant Mortality Rate, 1988, Measured in Deaths of infants under 1 year old per 1,000 Live Births (Excluding fetal mortality). Source: *Statistical Abstract of the United States*, 1993, p. 81 Table 112)

The Report: Protein Content of Breakfast Cereals

Common sense would expect data analysis to be cool and logical — with a clear plan and a clear execution. And, indeed, making data analysis look that way, at the end, shows good style: It leads to writing that is clear and to the point. But the straight forward logic of the public report you see in the evening news or read in a scientific journal has, often, little to do with the erratic and roundabout path by which "simple" truth is actually discovered. Data analysis has two phases — *doing* the analysis is one phase, *presenting* the analysis is another. And it would be hard to reverse engineer the rules of data analysis, hard to figure out how it was done, if all you were allowed to see were the final presentation. If nothing else, the presentation, slick, simple, and compelling, usually hides the number of hours of thinking that can lie behind a single graph, not to speak of hiding the bright ideas that the analyst followed to a conclusion only to find, at the conclusion, that the ideas led nowhere. I'm told that John von Neumann, an innovator in mathematics and computer science, once compared the discovery of a mathematical proof to the construction of a great cathedral: Like a cathedral, a mathematical proof is not complete until the scaffolding has been removed. So too with data analysis.

Even the methods I present, in a report, may be different from the methods I used in the act. For example, consider the simple "method" of computing an average: "Everyone knows" what an average is. And it is hard to present a report without writing down a few averages — average income, average age, and so forth. People feel comfortable with this sort of thing and, since you want them to understand your work, you have to accommodate their expectations. But the truth is that the median (the middle value in terms of size) and simple mathematics, which are rarely used in a presentation, are often used during the analysis — we often use one kind of technique when we are doing an analysis, and use other techniques when we are trying to get the idea across — one kind of technique when we are in the act, another when we communicate..

There is also a difference in attitude toward error, during and after —treating it one way in the report, but treating it another way, and much more aggressively during the pursuit. In the report, error is error, the luck of the draw, random deviation — a real world event subject to the mathematical uncertainties of probability. But in the act, during the pursuit, there is no such thing as error: Every strange event is searched for meaning: Why is this number low while that number is high — and neither of them is average? Is there information here? Is there pattern to the these events?

For homework, give me the “works”. We want to see how you handle the messy process of data analysis. And we want you to clean it up for a report.

Protein Content of Breakfast Cereals

The protein component of common breakfast cereals varies widely from essentially negligible quantities of protein to a large fraction of an adult's entire daily requirement. The purpose of this analysis was to document this variation and to examine the source of this variation. Results from examination of thirty common breakfast cereals confirm that protein content varies from negligible amounts of protein to a substantial fraction of the daily requirement for protein (30 40?? grams).

The analysis examined protein content among thirty common breakfast cereals in portions as they are commonly consumed, using data from Bowes and Church, Food Values of Portions Commonly Used, Harper and Row, 1975. The figure below demonstrates the range and variation of the portions: Typically these cereals provide two to three grams of protein per serving, about 5 to 10 percent??? of the minimum daily requirement. And careful selection of the cereal can find protein content that is two fold, three fold, or even greater than that of low protein cereals. At 10.2 grams of protein per serving one cereal, Gerber's High Protein Cereal, is in a class by itself while Quaker Puffed Rice, at 0.8 grams of protein provided only a negligible fraction of the daily requirement.

Context: Who, What, Where, ...

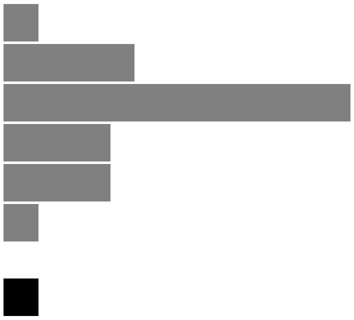
Newspaper style: Telegraph important results in the first paragraph. In professional papers, this would be an "abstract"

I looked up the minimum daily requirement to add context.

Source

Brief description: Typical, Extreme High, and Low

Grams of Protein per Serving	Number of Cereals
0 to 0.99 grams	1
1 to 1.99 grams	5
2 to 2.99 grams	14
3 to 3.99 grams	4
4 to 4.99 grams	4
5 to 5.99 grams	1
10 grams plus	1



Brief Pictorial Overview of the Data (Less detail than the Stem Leaf, but appropriate for this report

Change style to mark the break in the distribution of protein.

Figure 1

Range and Distribution of Grams of Protein per Serving for 29 Common Breakfast Cereals

While the data do not clearly indicate the principle ingredient for all of these cereals, fragmentary evidence suggests that the principle ingredient is the most reliable predictor of protein content. With considerable variation within groups, the averages range from 1.5 grams of protein for the four rice cereals to approximately 4 grams per serving for the oat and barley cereals, to 5.1 for the mixed corn/soy cereal, to a high of 10 grams of protein for the Gerbers soy cereal, summarizing the data shown in Figure 2.

Rice	**Corn**	**Bran (Grain not Specified)**	**Wheat**
Rice, Puffed, Quaker .8 Rice Cereal, Gerbers 1.5 Rice Krispies, Kellogg's 1.6 Rice Flakes 2.1	Corn Fetti, Post's 1.5 Corn Flakes 2.1	Bran Raisin, Kellogg's 1.8 Bran, Raisin, Post's 2.2 Bran Flakes, 40%, Post's 2.8 Bran Flakes 40% Kellogg's 2.9 Bran, All-Kellogg's 3.1	Wheat, Puffed Quaker 1.6 Wheat, Shredded 2.2 Wheaties, General Mills' 2.8 Wheat Chex, Ralston 2.8 Wheat Flakes, Quaker4.4
Average Grams per Serving **1.52**	**Average Grams per Serving** **1.8**	**Average Grams per Serving** **2.54**	**Average Grams per Serving** **2.76**
Oats	**Barley**	**Corn Soya**	**Soy**
Cheerios, General Mills' 3.4 Oatmeal 4.5			
Average Grams per Serving **3.95**	**Barley Cereal, Gerbers** **4.3**	**Corn Soya Shreds** **5.1**	**High Protein Cereal, Gerber's** **5.1**

Figure 2

Grams of Protein per Serving and Average Grams of Protein per Serving, organized by Principal Ingredient

Exercise:

Expand your analysis of fat in breakfast cereal, or the stem and leaf of the Dow industrials, or of U.S. infant mortality rates to a complete write up.

Dates:	**3/15/96**	**3/22/96**	***Change***
Dow Jones Industrial Average	**5,584.97**	**5636.64**	**51.67**
	Price per Share of Stock	Price per Share of Stock	
Alcoa Aluminum	$61.00	$62.38	$1.38
Allied Signal	$56.25	$57.13	$.88
American Express	$48.38	$48.75	$.37
AT&T	$61.38	$61.25	(-$0.13)
Bethlehem Steel	$13.75	$13.75	$0.00
Boeing	$80.88	$88.88	$0.00
Caterpillar	$72.00	$69.25	(-$2.75)
Chevron	$54.88	$55.25	$.37
Disney	$69.25	$64.88	($4.38)
duPont	$81.25	$83.00	$1.75
Kodak	$73.13	$73.13	$0.00
Exxon	$79.00	$81.38	$2.38
General Electric	$75.63	$78.38	$2.63
General Motors	$52.25	$53.38	$1.12
Goodyear	$51.25	$52.00	$ .75
IBM	$119.88	$114.25	(-$5.62)
International Paper	$39.25	$38.38	(-$0.88)
MacDonalds	$51.25	$50.75	(-$0.50)
Merck	$62.13	$63.50	$1.38
Minnesota Mining & Mfg	$63.50	$64.50	$1.00
Morgan	$80.00	$83.63	$3.63
Philip Morris	$95.38	$86.25	($9.25)
Proctor & Gamble	$83.13	$87.88	$4.75
Sears	$50.13	$51.00	$ .87
Texaco	$82.88	$84.75	$1.88
Union Carbide	$47.63	$48.25	$ .63
United Technologies	$111.00	$115.38	$4.38
Woolworth	$15.88	$15.75	(-$0.13)
Westinghouse	$19.00	$19.00	$0.00

Prices of “Dow Jones 30 Industrials”, March 15 – March 22, 1996

	Division and State	*Region*	*Total Infant Mortality Rate, 1988*
1		**U.S.**	**10.0**
2		**N.E.**	**8.1**
3	Maine	N.E.	7.9
4	New Hampshire	N.E.	8.3
5	Vermont	N.E.	6.8
6	Massachusetts	N.E.	7.9
7	Rhode Island	N.E.	8.2
8	Connecticut	N.E.	8.9
9		**M.A.**	**10.3**
10	New York	M.A.	10.8
11	New Jersey	M.A.	9.9
12	Pennsylvania	M.A.	9.9
13		**E.N.C.**	**10.5**
14	Ohio	E.N.C.	9.7
15	Indiana	E.N.C.	11.0
16	Illinois	E.N.C.	11.3
17	Michigan	E.N.C.	11.1
18	Wisconsin	E.N.C.	8.4
19		**W.N.C.**	**8.9**
20	Minnesota	W.N.C.	7.8
21	Iowa	W.N.C.	8.7
22	Missouri	W.N.C.	10.1
23	North Dakota	W.N.C.	10.5
24	South Dakota	W.N.C.	10.1
25	Nebraska	W.N.C.	9.0
26	Kansas	W.N.C.	8.0
27		**S.A.**	**11.6**
28	Delaware	S.A.	11.8
29	Maryland	S.A.	11.3
30	Dist Columbia	S.A.	23.2
31	Virginia	S.A.	10.4
32	West Virginia	S.A.	9.0
33	North Carolina	S.A.	12.5

34	South Carolina	S.A.	12.3
35	Georgia	S.A.	12.6
36	Florida	S.A.	10.6
37		**E.S.C.**	**11.4**
38	Kentucky	E.S.C.	10.7
39	Tennessee	E.S.C.	10.8
40	Alabama	E.S.C.	12.1
41	Mississippi	E.S.C.	12.3
42		**W.S.C.**	**9.4**
43	Arkansas	W.S.C.	10.7
44	Louisiana	W.S.C.	11.0
45	Oklahoma	W.S.C.	9.0
46	Texas	W.S.C.	9.0
47		**Mt.**	**9.2**
48	Montana	Mt.	8.7
49	Idaho	Mt.	8.8
50	Wyoming	Mt.	8.9
51	Colorado	Mt.	9.6
52	New Mexico	Mt.	10.0
53	Arizona	Mt.	9.7
54	Utah	Mt.	8.0
55	Nevada	Mt.	8.4
56		**Pac**	**8.6**
57	Washington	Pac	9.0
58	Oregon	Pac	8.6
59	California	Pac	8.6
60	Alaska	Pac	11.6
61	Hawaii	Pac	7.2

Total Infant Mortality Rate, 1988, Measured in Deaths of infants under 1 year old per 1,000 Live Births (Excluding fetal mortality). Source: *Statistical Abstract of the United States*, 1993, p. 81 Table 112)

Death in London

The Diseases, and Casualties this year being 1632.

Abortive, and Stilborn . .	445	Jaundies	43
Affrighted	1	Jawfaln	8
Aged	628	Impostume	74
Ague	43	Kil'd by several accidents	46
Apoplex, and Meagrom	17	King'sEvil	38
Bit with a mad dog	1	Lethargie	2
Bleeding	3	Livergrown	87
Bloody flux, scowring, and flux	348	Lunatique	5
Brused, Issues, sores, and ulcers	28	Made away themselves	15
Burnt, and Scalded	5	Measles	80
Burst, and Rupture	9	Murthered	7
Cancer, and Wolf	10	Over-laid, and staved at nurse	7
Canker	1	Palsie	25
Childbed	171	Piles	1
Chrisomes, and Infants	2268	Plague	8
Cold, and Cough	55	Planet	13
Colick, Stone, and Strangury	56	Pleurisie, and Spleen	36
Consumption	1797	Purples, and spotted Feaver	38
Convulsion	241	Quinsie	7
Cut of the Stone	5	Rising of the Lights	98
Dead in the street, and starved	6	Sciatica	1
Dropsie, and Swelling	267	Scurvey, and Itch	9
Drowned	34	Suddenly	62
Executed, and prest to death	18	Surfet	86
Falling Sickness	7	Swine Pox	6
Fever	1108	Teeth	470
Fistula	13	Thrush, and Sore mouth	40
Flocks, and small Pox	531	Tympany	13
French Pox	12	Tissick	34
Gangrene	5	Vomiting	34
Gout	4	Worms	27
Grief	11		

Christened	Males 4994 Females4590 In all 9584	Buried	Males 4932 Females 4603 In all 9535	Whereof, of the Plague 8

Increased in the Burials in the 122 Parishes, and at the Pest-

house this year	993
Decreased of the Plague in the 122 Parishes, and at the Pest-house this year	266

(From Newman, *The World of Mathematics*, Page 1425)

> “Moreover, if all these things were clearly, and truly known (which I have but guessed at) it would appear, how small a part of the People work upon necessary Labours, and Callings, viz. how many Women, and Children do just nothing, onely learning to spend what others get? how many … live by puzling poor people with unintelligible Notions in Divinity, and Philosophie, how many by perswading credulous, delicate, and Litigious Persons, that their Bodies, or Estates are out of Tune, and in danger. And on the other side, how few are employed in raising, and working necessary food and covering? and of the speculative men, how few do truly studie *Nature*, and *Things*? The more ingenious not advancing much further than to write, and speak wittily about these matters.
>
> — John Graunt, *Natural and Political Observations, London, 1662*

Death in London: Establishing Credibility — Who, What, Where, Why, When, and How?

More data: Figure 1 (facing) presents data on the causes of death in London, in 1632. That’s the target of my next data analysis and what I want to extract from the facts is information: I want to “know” what people died of in London three to four hundred years ago. And now let’s take it through the steps: For these data I need orientation, Who, What, Where, … and I need technique — this is more difficult than the data for breakfast cereals.

So, death in London, Who, What, Where: The number one question is probably “Why London, and why 1632?” I chose it because, according to *The World of Mathematics*” where I found these data, the original work by John Graunt, was one of the original uses of statistics for data analysis[1] It seemed appropriate to begin a discussion of data analysis by presenting one of the true beginnings of the profession. Even such "obvious" techniques as these began somewhere and had to be invented by a human mind like our own, like the mind of John Graunt.

Continuing, before the numbers, what are these data about and what's the context? I've already given you that: I've told you the data are about death and I've warned you that the data are very old, from another time. Still, I want to call attention to the obvious in order to warn you again: These are not numbers, these are data. Don't think about them as numbers, two plus two equals four. Think about life and death and old London — long before the invention of drugs that have kept me, for one, alive, before Pasteur with his understanding of bacteria, and only slightly after the invention of the microscope.

We also need to know what John Graunt was up to, it may have affected his results: Graunt seems to have known that he was an innovator, for whom the tabulation of diseases and casualties was just the beginning. His interest was in the state, England. He seems to have been interested in the efficient and balanced functioning of the state:

> I conclude, That a knowledge of all these particulars, and many more, whereat I have shot but at rovers, is necessary in order to good, certain, and easie Government, and even to balance Parties, and factions both in *Church* and *State*.
>
> — John Graunt, *op. cit*.

O.K., now — the numbers for death in London.

[1] From "Foundations of Vital Statistics", by John Graunt, reprinted in "The World of Mathematics", Volume 3, page 1421, published by Simon and Schuster, New York, 1956.

Death in London

The Diseases, and Casualties this year being 1632.

Abortive, and Stilborn . .	445
Affrighted	1
Aged	628
Ague	43
Apoplex, and Meagrom	17
Bit with a mad dog	1
Bleeding	3
Bloody flux, scowring, and flux	348
Brused, Issues, sores, and ulcers,	28
Burnt, and Scalded	5
Burst, and Rupture	9
Cancer, and Wolf	10
Canker	1
Childbed	171
Chrisomes, and Infants	2268
Cold, and Cough	55
Colick, Stone, and Strangury	56
Consumption	1797
Convulsion	241
Cut of the Stone	5
Dead in the street, and starved	6
Dropsie, and Swelling	267
Drowned	34
Executed, and prest to death	18
Falling Sickness	7
Fever	1108
Fistula	13
Flocks, and small Pox	531
French Pox	12
Gangrene	5
Gout	4
Grief	11
Jaundies	43
Jawfaln	8
Impostume	74
Kil'd by several accidents	46
King'sEvil	38
Lethargie	2
Livergrown	87
Lunatique	5
Made away themselves	15
Measles	80
Murthered	7
Over-laid, and staved at nurse	7
Palsie	25
Piles	1
Plague	8
Planet	13
Pleurisie, and Spleen	36
Purples, and spotted Feaver	38
Quinsie	7
Rising of the Lights	98
Sciatica	1
Scurvey, and Itch	9
Suddenly	62
Surfet	86
Swine Pox	6
Teeth	470
Thrush, and Sore mouth	40
Tympany	13
Tissick	34
Vomiting	34
Worms	27

Christened	Males	4994	Buried	Males	4932	Whereof, of the Plague 8
	Females	4590		Females	4603	
	In all	9584		In all	9535	

Increased in the Burials in the 122 Parishes, and at the Pest-house this year	993
Decreased of the Plague in the 122 Parishes, and at the Pest-house this year	266

(From Newman, *The World of Mathematics*, Page 1425)

Death in London:
Stem and Leaf — More Technique

The stem and leaf drawings used for the breakfast cereals required a little bit of experimentation with the stems and a lot of experimentation with the leaves, choosing the form that squeezed the most information out of the data. These new data present a pattern that is very common highly resistant to this stem and leaf technique. With these data no simple expansion or compression of the stems, or simple experimentation with the labels is going to get it really “right”. Instead, they require the use of several different scales, combined together in one drawing. As before, let me demonstrate the technique (and the problem) with raw numbers, plucked out of context, in order to work with technique. Here is the new set of numbers, shown in Figure 1.

445, 1, 628, 43, 17, 1, 3, 348, 28, 5, 9, 10, 1, 171, 2268, 55, 56, 1797, 241, 5, 6, 267, 34, 18, 7, 1108, 13, 531, 12, 5, 4, 11, 43, 8, 74, 46, 38, 2, 87, 5, 15, 80, 7, 7, 25, 8, 13, 36, 38, 7,98, 1, 9, 62, 86, 6, 470, 40, 13, 34, 1, 27.

Figure 1
More Numbers, for Practice

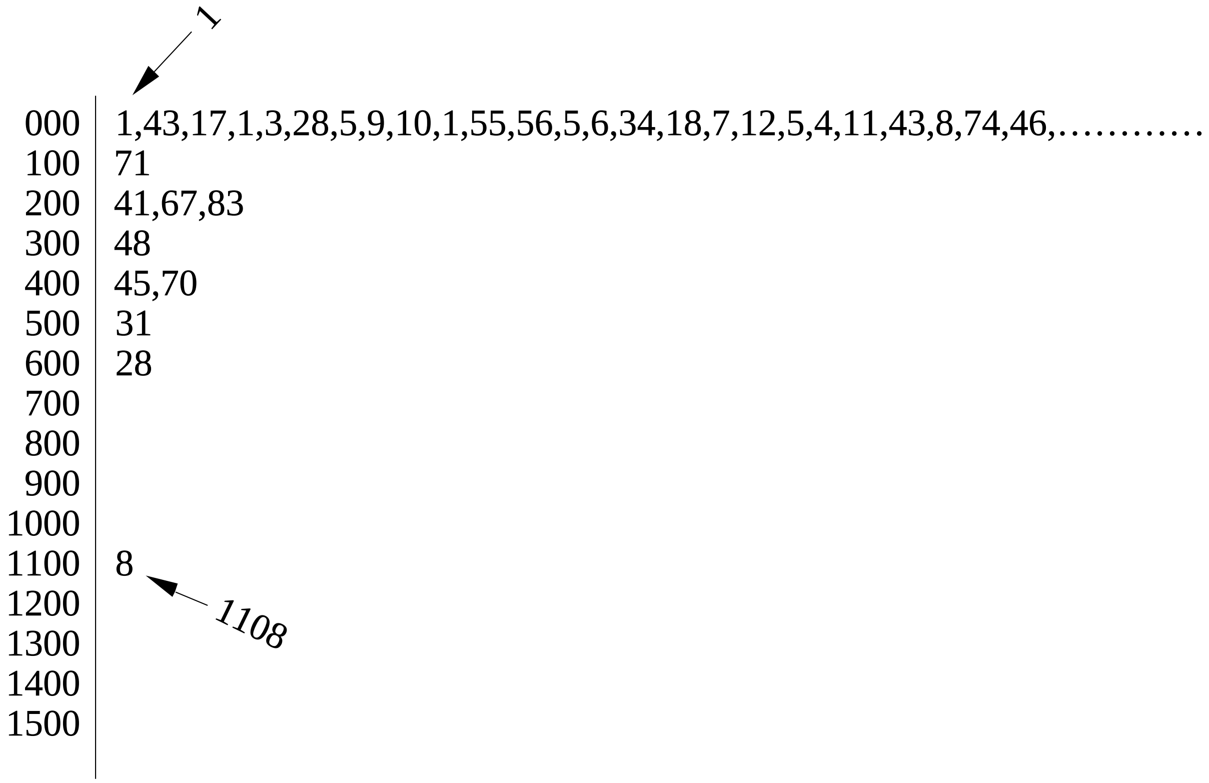
Suppose I try a straightforward approach, using 10’s for my stems: 0, 10, 20, 30, ..., Figure __.

```
             1
  00  |  11359156754325771871961
  10  |  708321533
  20  |  857
  30  |  48764
  40  |  3360
  50  |  56
  60  |  2
  70  |  4
  80  |  06
  90  |  3    93
 100  |
 110  |
 120  |
 130  |
 140  |
 150  |
  …   |
```

Figure 2
Stem and Leaf, by 10's

No, that's not going to work, I've got to extend the graph all the way up to 2268, which would fill up a couple of pages with this kind of stem.

Well, suppose I try 100's for my stems, compressing the graph, as in Figure 3. That's still going to have too many stems. Worse, it's still too big and it's bunching up a lot of data in a pile, undifferentiated.



000	1,43,17,1,3,28,5,9,10,1,55,56,5,6,34,18,7,12,5,4,11,43,8,74,46,…………
100	71
200	41,67,83
300	48
400	45,70
500	31
600	28
700	
800	
900	
1000	
1100	8
1200	
1300	
1400	
1500	

Figure 3
Stem and Leaf, by 100's

Stems of 200 would give me about the right number of stems, but most of the data would be in a pile: In fact, using 200's in Figure 3, and giving each leaf about the same amount of room on the page, I've only been able to squeeze about a quarter of the first set of leaves onto the page. Stems of 500 or 1000 would be worse.

0	1, 43, 17, 1, 3, 28, 5, 9,1 0, 1,171, 55, 56, 5, 6,………
200	348,241,267,283
400	445,531,470
600	628
1000	1108
1200	
1400	
1600	1797
1800	
2000	
2200	2268
2400	

Figure 3
Stem and Leaf, by 200's

There's the problem. And the solution for such a problem is not a great one, but it will do: The solution is to change scale one or more times, right in the middle of the graph. Here in Figure 4, for example, I've counted by 2's up to 10, then by 10's up to 100, then by 100's up to 1000, changing the values.

0	1, 1, 1, 1, 1
2	3, 4
4	5, 5, 5, 4, 5
6	6, 7, 7, 7, 7, 6
8	9, 8, 8, 9
10	17, 10, 18, 13, 12, 11, 15, 13, 13
20	28, 34, 38, 25, 36, 38, 34, 27
40	43, 55, 56, 43, 46, 40
60	74, 62
80	80, 98, 86
100	171
200	348,241,267
400	445,531,470
600	628
800	
1000	1797,1108
2000	2268,2837

Figure 4
Stem and Leaf, Variable Stems

Again I'll get a pile up — where I've changed the intervals. And I've actually used seven different scales: By 2's between 0 and 10, by 10's between 10 and 20, by 20's between 20 and 100, by 100's between 100 and 200, by 200's between 200 and 1,000, and by 1,000's between 1,000 and 2,000. That's a rather unpleasant and unsatisfactory solution. And what you'll see later, is that the unpleasant and unsatisfactory feeling you should be getting from these data is itself a clue — to a better solution, later. But for now —

Death In London: The Work

Back to the data — in context. Back to death in London. Now I, for one, am unfamiliar with the diseases listed, including “Rising of the Lights” and “Kings Evil”, and there is a tendency to chuckle a little at the peculiarity of these causes of death. I want you to use that chuckle and the peculiarity of the stem and leaf — and every other clue you can muster to make sense of these data: Observe yourself and your reactions to the data. What’s going on? Keeping that in mind, let’s begin getting a feel for the data.

The first thing I do is simply look at the page. I read some of the labels. Ah, this is unfamiliar turf. I look at the labels on the whole page, “Diseases and Casualties this year being 1632”. Ah, mortality statistics filtered through the lens of another culture, before the biology and medicine of our own era. These are not “our” categories, nor our definition of disease. What do I expect from these data? I expect to be mystified, confused by descriptions of disease from a “pre-scientific” era. If anything, I remember something about the plague. Is that relevant? No, only 8 deaths from plague. So I’m expecting confusion, causes as meaningless to me, three hundred years later, as arterial sclerosis would have been to them, three hundred years ago.

Looking at the numbers: They start out with a “445” and continue with a “1” — some of these categories are big, some are small. In fact, they range from over 2,000 down to 1. Looking at the bottom of the page there is some special attention to the plague, but the number dying from it is small. How big is the population experiencing this death. I don’t know, but I can guess: The number of people Christened is about ten thousand as is the number buried. If a modern birth or death rate is about 1 or 2 %, then I can guess that London of 1632 already had a population somewhere between one hundred thousand and one million. That's a rough estimate of the order of magnitude for this population. Let's look at more numbers — and now put some order into the search.

Looking at the detail, I'll use the Stem and Leaf to re-arrange these "diseases" according to size, rare to frequent, and take a look. Even if I had a computer available, one that would "do it right" the first time, I would begin with the stem and leaf, by hand, because it will force me to read labels like "Abortive and Stilborn" and the names of each of these other categories: Doing the stem and leaf, by hand, will rivet my attention on these data. I want about ten stems, I want the definitions of these stems to be easy (so my attention does not get diverted), and so I'll try "stems" beginning at 0, and going up, beginning at 100, and going up, beginning at 200, and going up, ...

0	
100	
200	
300	
400	
500	
600	
700	
800	

. . .

No, that's getting me a lot of categories (as I observed earlier when I treated these numbers as raw numbers). I've got a long way to go before I get to 2,000 (for Chrisomes and Infants). Observing this little numerical difficulty, deciding for myself (without the aid of a computer) how to organize these stems, and finding some difficulty, — I observe my difficulty and learn something about the data: Some of these things are really big, really "off scale" as compared to others. Note this: The "data analysis" lies in observing and learning, not in finishing another "stem and leaf".

So, I'm going to use two scales, perhaps 0 to 1,000 plus a stem for and "Other." So I begin to write

0	Affright, ague, apoplex, bit, bleeding, brused
100	
200	
300	Bloody
400	Abort
500	
600	Aged
700	
800	

. . .

Pausing, with nine leaves, "everything" (six of the first nine) is going in to the first category, failing to differentiate among these things. Let me try again, expanding the stem and leaf.

0	Affright, bit with mad dog, bleeding
10	Apoplex, and Meagrom
20	
30	
40	Ague
50	
60	
70	
80	

. . .

100	
200	
300	Bloody flux
400	Abort
500	
600	Aged
700	
800	

. . .

1000	
2000	

O.K., good enough for now. I'll continue using three ranges. And of course the breaks in ranges were chosen for convenience, not because the data break at precisely those points. Continuing:

0	Affright; bit with mad dog; bleeding; burnt and Scalded; Burst & Rupture; Canker;
10	Apoplex, and Meagrom; Cancer and Wolf
20	Brused issues
30	
40	Ague
50	Cold and Cough; Colick Stone
60	
70	
80	

. . .

100	Childbed
200	Convulsion
300	Bloody flux
400	Abort
500	
600	Aged
700	
800	

. . .

1000	Consumption
2000	Chrisomes and Infants;

Nineteen leaves — getting messy: The data are piling up in the smallest stem. I'm thinking, "Perhaps I could clean up my technique by introducing yet another range." But for now, I'll just let them pile up and get on with it — I want to look at the data. Beside which, my mind is beginning to focus on these things — which means that the stem and leaf is doing its job: I'm more interested in "Chrisomes", than I am in "bit with mad dog", and I don't want to spend too much time holding myself back with technicalities: I'm trying to learn something: I'm thinking "What are 'chrisomes' ?" and the stem and leaf has me building up a list of names to check as soon as I can find a dictionary that's likely to include such things. In fact, I'm thinking, "I'm not going to go on automatic and finish this thing, as if I were a computer: I'm going to focus on the big stuff, like Chrisomes. That's where I'll learn something," So, continuing again, but limiting myself to the big stuff:

. . .

100	Childbed
200	Convulsion; Dropsie and Swelling
300	Bloody flux
400	Abort; teeth
500	Flocks and small pox
600	Aged
700	
800	

. . .

1000	Consumption; Fever
2000	Chrisomes and Infants;

That's simple. Call it "selective" — having ignored most of the data. But that's just being focused. I've decided that the information lies in the big stuff and that's what I'll look at. And what do I see? Well, in reverse order, I still don't know what Chrisomes are, but they have something to do with infants. "Consumption?" Ah, I remember that one: tuberculosis, associated with crowded conditions in the absence of things like clean air and clean water, a public health problem. "Fever?", too general. "Abortive", let's look back for the full label: "Abortive and Stilborn". Ah, we're dealing with childbirth again. "Bloody flux", well— considering that I've already got two categories related to childbirth, I can guess about that one ... and there's "childbed" in a nearby stem. "Teeth"? I'm beginning to guess. I'll bet we're talking about infections, (using a modern definition). And I'm beginning to get a picture: The people at risk are newborns, fertile and child bearing women, and people with some sort of susceptibility to infection: And, more generally, I'm thinking that this London was a filthy mess in which, if you got an infection, you died.

Now, to the dictionary: I still want to know about "Chrisomes" [ ], O.E.D, — consistent with what I guessed and my hunches about London in 1632: Expose your blood to the environment, expose your mucus membranes to the environment (colick, consumption, pleurisie, sore mouth,) and you die, or you get sick (fever) and then you die.

That is what data analysis is about: Wringing the data for information about the world behind the data, for reasonable hunches and a direction that guides the next step in my study. Right or wrong, so far, it's got me thinking and building hypotheses about this London of 1632.

Death in London: The Report

Use photo copy

The Diseases, and Casualties this year being 1632.

Tabulation by John Graunt, reproduced from James R. Newman's *The World of Mathematics, Volume 3*, page 1421, published by Simon and Schuster, New York 1956.

Figure 1
Causes of Death in London, 1632

Life in London, three hundred years ago, was dirty and short. That was the picture disclosed in one of the first compilations of vital statistics, published by John Graunt in 1662.

His data described 10,000 deaths from many causes, most of them unknown to the modern vocabulary. But a few causes account for the vast majority of mortality. Most striking, comparing the number of christenings to the number of deaths among the young, infant mortality probably exceeded twenty-five percent (comparing 2,300 deaths from "Chrisomes and Infants" to 9,600 christenings).

Exotic causes that are popular in the current image of historical London accounted for very few of the actual deaths. In this year, for

> example, plague accounted for only 8 deaths, By contrast, the major causes of death are appalling familiar to us, even three hundred and fifty years later: Adults die from infections and communicable diseases of which one example, tuberculosis (consumption), accounted for approximately 20% of the mortality. Other categories suggesting infection or contagious disease account for another third of all deaths — including “Fever”, “Flocks and Small Pox”, “Brused …”, “Cold…”, “Dropsie …”, “Convulsion”, “Childbed”, “Bloody flux ...”, and “Teeth”. Adding it up, death related to childbirth, infections, and communicable diseases accounted for approximately eighty percent of mortality.
>
> The data were tabulated by John Graunt from weekly Bills of Mortality and published in 1662. Causes of death were report by attending physicians and bystanders. The report includes detail with respect to sources of data and comments on likely sources of error. Excerpts from Graunt’s report are available in reprint as “Foundations of Vital Statistics”, in James Newman’s *The World of Mathematics*,, noted above. Graunt’s stated purpose was to distinguish fact from fiction, to compare true causes of death to those, such as plague, that commanded public attention and for the general purpose of increasing the welfare of the state. Grant’s report included extended comments on the quality of the data and likely sources of error.

There’s my report, less than five hundred words — probably a bit longer than I would ordinarily write because I have students looking over my shoulder. Now, how did I go from Graunt’s table to my summary? The key, and the focus of this discussion of method was the “Stem and Leaf” diagram worked-out earlier. But I haven’t even presented it here in my report. It was essential to the process, but once I focused on the few causes of the overwhelming number of deaths, I left it behind. Little would have been added to the reader’s knowledge of death in London if I had recapitulated my steps in the final report. That’s hard for a writer: All that work — hidden. No one will know to applaud my diligence. But that’ the difference between writing a report about *London* and writing a report about *yourself* — showing your reader how much hard work went into the report. The reader is interested in London. So, spare the reader — and keep the focus.

Exercise: Prepare a report on causes of death in the United States, 1990, using these U.S. Census data.

From the *Statistical Abstract of the United States, 1992,* Table Number 114: Deaths and Death Rates, by Selected Causes, 1990 (1990 estimates are preliminary data from a 10 percent sample including deaths of non-residents.

Cause of Death	**Deaths (1,000), 1990**
All Causes	**2,162.0**
Major cariovascular diseases	920.4
Diseases of heart	725.0
Percent of total	33.5
Rheumatic fever and rheumatic heart disease	6.3
Hypertensive heart disease[1]	23.6
Ischemic heart disease	489.3
Other diseases of endocardium	12.3
All other forms of heart disease	193.5
Hypertension[1]	9.2
Cerebrovascular diseases	145.3
Atherosclerosis	16.5
Other	24.4
Malignancies[2]	506.0
Percent of total	23.4
Of respiratory and intrathoracic organs	143.8
Of digestive organs and peritoneum	121.3
Of genital organs	58.0
Of breast	45.1
Of urinary organs	20.4
Leukemia	18.7
Accidents and adverse effects	93.6
Motor vehicle	47.9
All other	45.7
Chronic obstructive pulmonary diseases and allied conditions	89.0
Bronchitis, chronic and unspecified	3.4
Emphysema	16.5
Asthma	4.6

[1] With or without renal desease

[2] Includes other types of malignancies not shown separately

Other	64.6
Pneumonia and influenza	78.6
Pneumonia	76.7
Influenza	1.9
Diabetes melitus	48.8
Suicide	30.8
Chronic liver disease and cirrhosis	25.6
Other infective and parasitic diseases	32.2
Human immunodeficiency virus (HIV) infections (AIDS)	24.1
Homicide and legal intervention	25.7
Nephritis, nephrotic syndrome, and nephrosis	20.9
Septicemia	19.8
Certain conditions originating in the perinatal period	17.5
Congenital anomalies	13.4
Benign neoplasms[3]	7.0
Ulcer of stomach and duodenum	6.2
Hernia of abdominal cavity and intestinal obstruction[4]	5.6
Anemias	4.2
Cholelithiasis and other disorders of gall bladder	3.0
Nutritional deficiencies	3.1
Tuberculosis	1.8
Infectionss of kidney	1.1
Viral hepatitis	1.7
Menengitis	1.2
Acute bronchitis and bronciolitis	0.6
Hyperplasia of prostate	0.3
Symptoms, signs, and ill-defined conditions	26.3
All other causes	174.1

Deaths classified according to ningth revision of *International Classification of Diseases*. Original source: U.S. National Center fo Health Statistics, *Vital Statistics of the United States*, annual; Monthly Vital Statistics Report and unpublished data.

[3] Includes neoplasms of unspecified nature and carcinoma in situ.

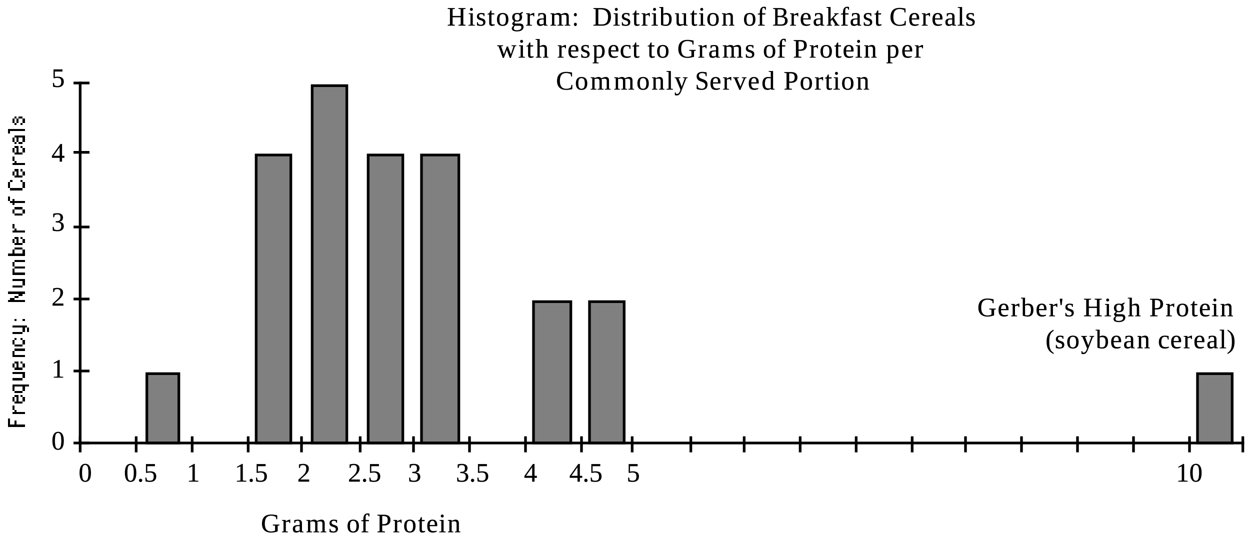
[4] Without mention of hernia

There are blanks in here, presumably histograms that need to be computed and drawn

Also extract each data set into a separate file (duplicating what is in the text). Put into the folder of data.

Histograms

The most widely used graphical device for the *formal* display of a distribution the histogram. For example, Figure 1 is a histogram of distribution of protein count from breakfast cereals.



The histogram is a stylized version of the stem and leaf — cleaned-up for public inspection, with much of the information removed. It’s the kind of thing you might use after the analysis is completed — when you are constructing a “pretty” final report — or use when the number of objects involved is massive as, for example, the distribution of family incomes for 100-million families, or when you use a computer program that takes some care with its graphics.

My only objection to these things is that, because they are "clean," and often computer drawn, the impression is created that they are somehow better or, at least, more scientific than the stem and leaf — when exactly the opposite is the case: That "mess" in the stem and leaf is information, information that is not available in the histogram. And for that reason, the histogram is not the kind of thing that a real human being does early, as a first resort — when you are in hot pursuit of information. For that, by hand, most people would use tallies, ~~////~~, or digits or labels, as I've done in the stem and leaf drawings.

That being said, and the warning having been sounded, histograms are often used for income distributions; they play an important role in numerical work on ecology,[1] and they are useful for visually contrasting two different distributions — where they serve much the same function as a stem and leaf diagram.

I want to examine the construction of a histogram in painful detail because there are a few tricks to it, a few places where its construction is different from the construction of the stem and leaf. In particular, you should focus on two questions: First, what is the vertical axis of the histogram, the "height?" The height is not always proportional to the simple count (as it was in the stem and leaf). And, second, what is the area of the histogram, the "area" under the curve? (Usually, these things are simply unlabelled — the computer doesn't know what they are. But you have to know what they are in order to build a histogram for yourself, and you have to be able to build one for yourself in order to be sure you understand it.)

For inspection, here is a set of histograms: A histogram of gross national products (by nation), a histogram of numbers of animals per species within a ____, and a pair of histograms showing the family income by "race" in the United States.

[1] See *Ecological Diversity and Its Measurement* by Anne E. Magurran, Princeton University Press, 1988.

Use the SocPol data for Gross National Products or new data from the World Bank.

Use the biodiversity data from Magurran

Use the U.S. census, family incomes

Histograms look simple, and *are* simple for data like the distribution of protein content of breakfast cereals. But when you get to data like the income distributions, you have to be clear about the construction rules. The problem is that the widths of the categories, $1,000 at one end (e.g., from $1,000 to $1,999) and $4,000 at the other (from 10,000 to 14,999). To get it right, let me look at the construction of this histogram in painful detail.

The key point to remember in constructing a histogram is that the shaded area is proportional to the number of “things” whose distribution is being described. Knowing the meaning of the area, answers both questions at once: If the “things” are thirty breakfast cereals, then the shaded area is proportional to thirty breakfast cereals. If the “things” are 100 percent of the income earning families, then the shaded area is proportional to 100 percent.

For the first detailed example, I’m going to build the histogram for U.S. family income. Figure _, the first two columns on the left, shows the data exactly as they come from the book (The U. S. Book of Facts Statistics, and Information, 1972, page 316.)

Note: All I have handy is old data, for some reason. Revise to something more recent.

Money Income Percent Distribution of Families by Income Level, Table 500, page 316, U.S. Book of Facts Statistics and Information, 1972.

Income Level	Population in Percent %
Under $1,000	1.6%
$1,000 to $1,999	3.1%
$2,000 to $2,999	4.6%
$3,000 to $3,999	5.3%
$4,000 to $4,999	5.4%
$5,000 to $5,999	5.9%
$6,000 to $6,999	6.4%
$7,000 to $9,999	21.7%

Income Interval in dollars $	Height in percent per dollar %/$	Height in simple numbers (without units) used for drawing the histogram
~$1,000	~.0016 %/$	16
$1,000	.0031 %/$	31
$1,000	.0046 %/$	46
$1,000	.0053 %/$	53
$1,000	.0054 %/$	54
$1,000	.0059 %/$	59
$1,000	.0064 %/$	64
$3,000	.0072 %/$	72

$10,000 to $14,999	26.7%	$5,000	.0053 % / $	53
$15,000 or over	19.2%	Ill-defined guess $100,000	Ill-defined guess .00019 % / $	02

Now, to construct the histogram. First I get out my graph paper and mark off the income levels, left to right, marking $1,000, $2,000, and so forth. These marks give me the left and right boundaries of each piece of the histogram. But note, the intervals are not equal in size.

supply graph in progress.

Now, all I have to do is supply a height for the shaded area over each of the intervals, a height for the area between 0 and $1,000, a height for the area between $1,000 and $2,000, and so forth. Here's where I fill in the last three columns of the table:

For the first area, here's what I know: I know that the area is 1.6% and I know that the width of the interval is approximately $1,000. And I also know, from simple geometry that

Area = Width times Height

So, what's the height? Simple: If area equals width times height then height equals area divided by width

Height = Area / Width

Using the equation for this particular problem, the first equation, area equals width times height means

1.6% = $1,000 times Height

And so, the height is

$$\text{Height} = \frac{1.6\%}{\$1{,}000}$$

which is

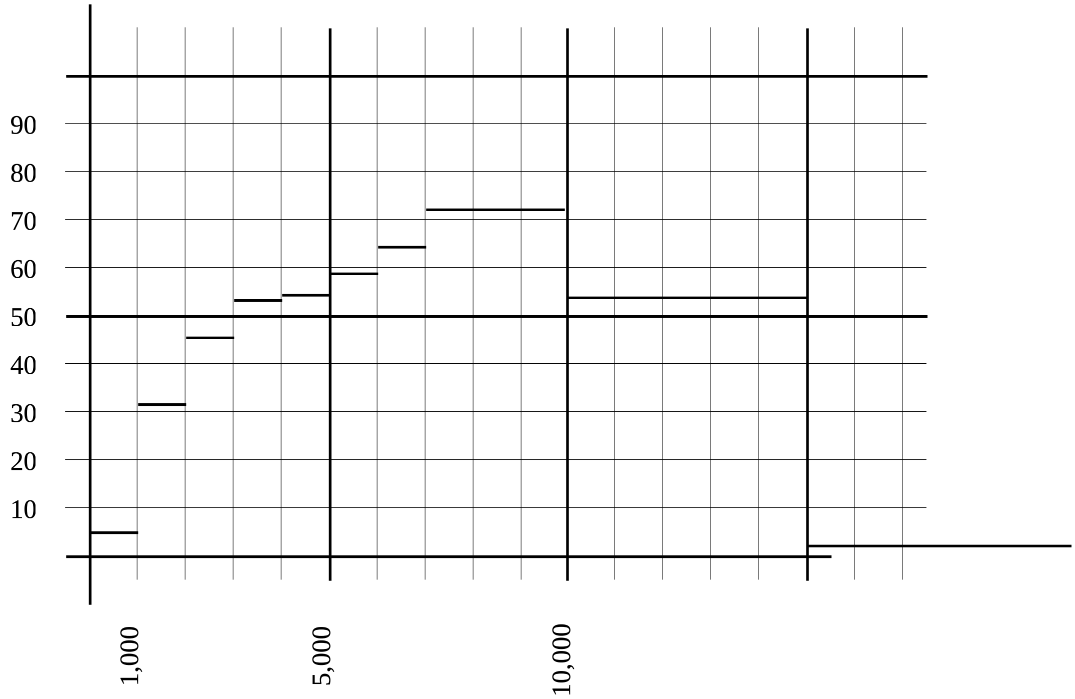
.0016 ***percent per dollar***

That is the height of this first part of the histogram, and those are the units on the vertical axis: The height measures percent of the population per dollar of family income.

Incomplete graph, with one piece of the histogram, and the vertical axis labeled in units from 0 to 100 percent of the population per dollar of family income.

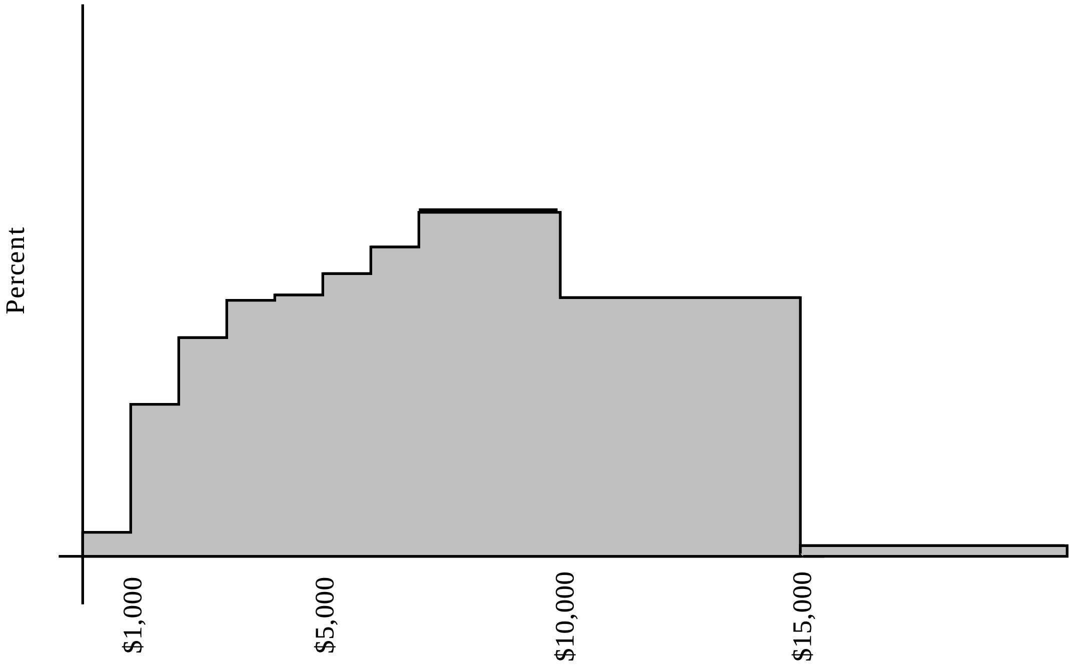
The thing you have to compute in order to draw the histogram is the height even though the data that you get describe the base and the area. So you have to compute the height which is, in this case, percent of the population per dollar of family income. The thing that you have to get right is the height, while the data describe the base and the area. So we compute the height that makes the area right.

Having computed the height, I get out my graph paper and translate the numbers onto a grid.



And, then I begin to clean things up. At the low end, I'm not really sure that $0 is really the bottom on income — you can do worse than break even. But I'll assume that "0" is the left end of the first interval. At the high end, for the interval above $15,000, I have a more difficult problem. In fact, it is an insoluble problem and there's not much I can do about it: Fact is, I don't know what to use for the highest income, which makes the width of the last piece of the graph entirely arbitrary. And since the width is arbitrary, the height of this piece of the histogram is also arbitrary. You could reasonably ask for better data: Lumping everyone above $15,000 dollars is wiping out an awful lot of detail. But even with better data, there's always going to be a last category and often, with income data, there's no answer. I've arbitrarily chosen $100,000 as my top, and forged ahead with my arbitrary choice. And here is the result

(fix the label on the vertical axis -- in fact, do the whole thing over again on more recent data)



To take a look at a more “classical” shape, consider this histogram of total population of nations, using 1975 data from the *World Handbook of Social and Political Indicators*

(check World Bank for more recent data. Show both the chart and the histogram.)

Homework:

Family Income 1991, General Social Survey, National Opinion Research Center, University of Chicago

Income Level	Number of Families (Sample size = 1517) families
Under $1,000	11
$1,000 to $2,999	26
$3,000 to $3,999	31
$4,000 to $4,999	35
$5,000 to $5,999	39
$6,000 to $6,999	29
$7,000 to $7,999	23
$8,000 to $9,999	38
$10,000 to $12,499	76
$12,500 to $14,999	82
$15,000 to $17,499	97
$17,500 to $19,999	60

Income Interval in dollars $	Height in number per dollar families / $	Height (for use on graph)

$20,000 to $22,499	60				
$$22,500 to $24,999	68				
$25,000 to $29,999	112				
$30,000 to $34,999	94				
$35,000 to $39,999	86				
$40,000 to $49,999	149				
$50,000 to $59,999	86				
$60,000 to $74,999	86				
$75,000 and higher	80				
Refused	85				
Don't Know	47				
No Answer	17				

Household size: Number of Household Members (General Social Survey, 1991, Var “HOMPOP, #33”

Household Size	Number of Respondents (n=1517)
1	377
2	476
3	275
4	241
5	98
6	29
7	14
8	2
9	2
10	2
No Answer	1

Respondent's Education General Social Survey, 190, Variable "EDUC", #15

Grade or Years	Number of Respondents (n=1517)
Schooling	2
1st grade	0
2nd grade	0
3rd grade	5
4th grade	5
5th grade	6
6th grade	12
7th grade	25
8th grade	68
9th grade	56
10th grade	73
11th grade	85
12th grade	461
1 year of Coll	130
2 years	175
3 years	73
4 years	194
5 years	43
6 years	45
7 years	22
8 years	30
Don't Know	0
No Answer	0

Respondents Age VARIABLE: AGE #12

Age	Number of Respondents (n=1517)
10 - 19	12
20 - 29	293
30 - 39	382
40 - 49	280
50 - 59	165
60 - 69	171

70 - 79	148
80 or over	63
No answer, Don't know	3

Variable SIBS #10: How many brother and sisters did you have? Please count those born alive, but no longer living, as well as those alive now. Also include stepbrothers and stepsisters, and children adopted by your parents.

Number of Siblings	Number of Respondents (n=1517)
0	74
1	236
2	276
3	236
4	209
5	118
6	80
7	81
8	58
9	47
10	34
11	22
12	11
13	9
14	5
15	3
16	1
17	2
18	1
19	0
20	0
21	1
22	0
23	0
24	0
26	1
Don't know	4
No Answer	8

VARIABLE: HRS1, If working, full or part time: How many hours did you work last week, at alljobs

Number of Hours	Number of Respondents (n=1517)
0 - 09 hours	25
10 - 19 hours	55
20 - 29 hours	75
30 - 39 hours	117
40 - 49 hours	397
50 - 59 hours	114
60 - 69 hours	65
70 - 79 hours	17
80 or more	18
No ans, don't know	1
Not Applicable	633

Ed: Get each of the above separated on “race”, gender, and perhaps, as a table, age, educ, income.

The SRC data for education is better, showing military, trade school, etc.

Description:

Numbers for the Variation, Numbers for the Average

Data analysis is cumulative: *If* the data pass the first test that establishes whether or not it is worthwhile to continue, and *if* the stem and leaf begins to make sense of the data *then* it may be time to introduce numbers that summarize what has been observed in the stem and leaf.

The arithmetic by which we summarize data is very simple: It takes no more than a few minutes to master the arithmetic of an average. If anything, the arithmetic is too simple, causing analysts to compute the numbers, report them, and quickly move forward to something more mathematically difficulty and, presumably, more sophisticated. But my real job in these pages is to teach data analysis, not arithmetic, and to accomplish that I have to rivet your attention on two issues.

The first issue is variation: The whole idea of "average," assumes variation. If the income of the *American family* is $30,000, every family, that means one thing. If the *average* income of American families is $30,000, that means something else: It means some families have less than $30,000, some families have more than $30,000. And, in the middle, very few families will have exactly $30,000.00 — to the penny. Incomes vary. Do "typical" incomes vary between $29,995 and $30,005, or do typical incomes vary between $10,000 and $100,000: As data, as a message, the variation makes a big difference for any analysis of family incomes in the United States.

The second issue is strategy. I could simply list a couple of formulas with instructions for their use. And, in fact, I will do just that after I have established a context. But why use one technique or another? Is there a logic? Is there a consistency that logically binds one technique to another. There is, of course. And that, the ;underlying logic of the statistical tools is based on strategy.

One's bed is "unreal." The Idea of the bed, existing eternally in some distant empyrean, is the true reality. …… Any bright Athenian could have made the obvious objection to this stratospheric nonsense.

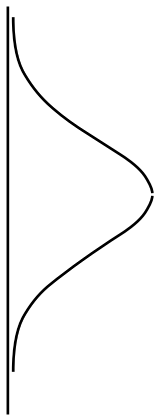
I. F. Stone, *The Trial of Socrates*, p. 73.

deviate: … to turn aside (from a course direction, standard, doctrine, etc.) …

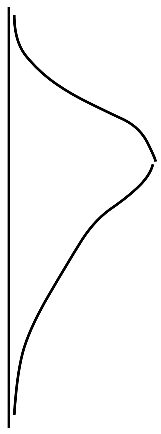
Webster's New World Dictionary of the American Language.

Things Vary

The moral of the story that follows is simple. The moral is "Things vary." There is a pleasant feeling of certainty to wrapping up your data with one definitive number: "The average income is $50,000." "The average number of people per physician is 512." That's it, a neat clean description of reality. But, reality is not that neat. Numbers vary. Data vary. It is almost guaranteed that the data are a lot messier than anything that could be reported by a single number. Look at the stem and leaf for a set of data and you will see a shape: Generally, you will see something with a central "hump" that we can think of, roughly, as the center of the distribution. Schematically, which is to say smoothing over the roughness of real data, the hump may be in the middle of the range or closer to one end, like



or like



The average, either the median or the mean, puts a number on the location of the center: “The median income of this population is $50,000.” “The mean income of this population is $60,000.” But we can do better by thinking of the “center” as a range of data, a hump that may be tightly wrapped around the average, or loosely spread. Generically, this tightness of looseness around the center is called “variation” and it makes a great difference to your analysis of data.

Measure your height, measure your weight, measure these things again and again, and the answers will vary. This is not because you used a bad yardstick or a cheap scale. It is because things vary.

I don’t understand just why variation, not constants, are the reality of our experience, but the idea that there is such a thing as *the answer* is just that: an idea. By contrast, what we know for sure, the evidence before our eyes, our experience, is variation.

For example, consider the weight of the standard 10 gram weight that resides at the United States Bureau of Standards. What does it weigh? That should be simple enough, a question shorn of the usual subtleties plaguing measurement in the social and physical sciences. Or is it?

The US 10 gram weight is not “the” 10 gram weight. The U.S. standard is a copy of the International standard 10 gram weight in France. It is an imperfect copy, a little bit lighter than the original. But what does it weight?

Table 1 records observations made by the U.S. Bureau of Standards itself (presented by David Freedman in *Statistics*[1]). Imagine the resources of the Bureau of Standards — fully capable of commanding whatever resources it takes to do the job — which in this case is to measure the weight of this little piece of metal.[2] And the answer? It varies. The figure shows one hundred observations (and nine different answers).

How much does it weigh? The measurements exhibit the distribution of values shown in Figure 1. It is *about* 0.4 milligrams light. *Most* of the measurement lie in a range between 9.99950 and 9.99960 grams, a range of uncertainty equivalent in weight to the weight of about half a centimeter (about three-sixteenths of an inch) of human hair.

[1] Statistics, Second Edition, by Freedman, Pisani, Purves, and Adhikari, Norton, 1991, page 93.

[2] The US 10 gram weight is not “the” 10 gram weight. The U.S. standard is a copy of the International standard 10 gram weight in France. It is an imperfect copy, a little bit lighter than the original. But what does it weight?

Item	Weight in Grams	Difference: Weight in Grams Minus 10 Grams
1	9.999591	-0.000409
2	9.999600	-0.000400
3	9.999594	-0.000406
4	9.999601	-0.000399
5	9.999598	-0.000402
6	9.999594	-0.000406
7	9.999599	-0.000401
8	9.999597	-0.000403
9	9.999599	-0.000401
10	9.999597	-0.000403
11	9.999602	-0.000398
12	9.999597	-0.000403
13	9.999593	-0.000407
14	9.999598	-0.000402
15	9.999599	-0.000401
16	9.999601	-0.000399
17	9.999600	-0.000400
18	9.999599	-0.000401
19	9.999595	-0.000405
20	9.999598	-0.000402
21	9.999592	-0.000408
22	9.999601	-0.000399
23	9.999601	-0.000399
24	9.999598	-0.000402
25	9.999601	-0.000399
26	9.999603	-0.000397
27	9.999593	-0.000407
28	9.999599	-0.000401
29	9.999601	-0.000399
30	9.999599	-0.000401
31	9.999597	-0.000403
32	9.999600	-0.000400
33	9.999590	-0.000410
34	9.999599	-0.000401
35	9.999593	-0.000407
36	9.999577	-0.000423
37	9.999594	-0.000406
38	9.999594	-0.000406
39	9.999598	-0.000402
40	9.999595	-0.000405
41	9.999595	-0.000405
42	9.999591	-0.000409
43	9.999601	-0.000399
44	9.999598	-0.000402
45	9.999593	-0.000407
46	9.999594	-0.000406
47	9.999587	-0.000413
48	9.999591	-0.000409
49	9.999596	-0.000404
50	9.999598	-0.000402
51	9.999596	-0.000404
52	9.999594	-0.000406
53	9.999593	-0.000407
54	9.999595	-0.000405
55	9.999589	-0.000411
56	9.999590	-0.000410
57	9.999590	-0.000410
58	9.999590	-0.000410
59	9.999599	-0.000401
60	9.999598	-0.000402
61	9.999596	-0.000404
62	9.999595	-0.000405
63	9.999608	-0.000392
64	9.999593	-0.000407
65	9.999594	-0.000406
66	9.999596	-0.000404
67	9.999597	-0.000403
68	9.999592	-0.000408
69	9.999596	-0.000404
70	9.999593	-0.000407
71	9.999588	-0.000412
72	9.999594	-0.000406
73	9.999591	-0.000409
74	9.999600	-0.000400
75	9.999592	-0.000408
76	9.999596	-0.000404
77	9.999599	-0.000401
78	9.999596	-0.000404
79	9.999592	-0.000408
80	9.999594	-0.000406
81	9.999592	-0.000408
82	9.999594	-0.000406
83	9.999599	-0.000401
84	9.999588	-0.000412
85	9.999607	-0.000393
86	9.999563	-0.000437
87	9.999582	-0.000418
88	9.999585	-0.000415
89	9.999596	-0.000404
90	9.999599	-0.000401
91	9.999599	-0.000401
92	9.999593	-0.000407
93	9.999588	-0.000412
94	9.999625	-0.000375
95	9.999591	-0.000409
96	9.999594	-0.000406
97	9.999602	-0.000398
98	9.999594	-0.000406
99	9.999597	-0.000403
100	9.999596	-0.000404

Figure 1 (Facing)

One hundred measurements of the weight of the U.S. Ten Gram Weight. (Statistics, Second Edition, by Freedman, Pisani, Purves, and Adhikari, 1991, Norton, page 93.)

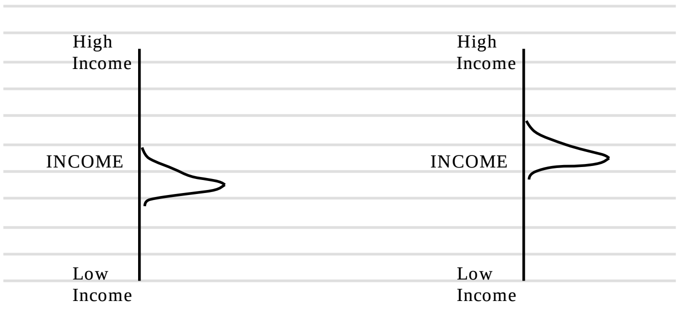
<table>
<tr><th>Weight Interval
in Grams</th><th>Distribution</th></tr>
<tr><td>9.999625</td><td>|</td></tr>
<tr><td>9.999620</td><td></td></tr>
<tr><td>9.999615</td><td></td></tr>
<tr><td>9.999610</td><td></td></tr>
<tr><td>9.999605</td><td>| |</td></tr>
<tr><td>9.999600</td><td>| | | | | | | | | | | | | |</td></tr>
<tr><td>9.999595</td><td>| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |</td></tr>
<tr><td>9.999590</td><td>| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |</td></tr>
<tr><td>9.999585</td><td>| | | | | |</td></tr>
<tr><td>9.999580</td><td>|</td></tr>
<tr><td>9.999575</td><td>|</td></tr>
<tr><td>9.999570</td><td></td></tr>
<tr><td>>=9.999565 to <9.999570</td><td></td></tr>
<tr><td>>=9.999560 to <9.999565</td><td>|</td></tr>
</table>

Table 1
Histogram of One Hundred Measurements of the 10 Gram Weight.

So here is our task: To summarize the facts displayed by a Stem and Leaf, or by a histogram, we need a number to represent the center of the variation, and we need a number (or numbers) to represent the variation.

Implicitly, you are already paying attention to variation: For example, if you had found that the variation of protein content among breakfast cereals was small, with protein content lying between a low of 2.3 grams of protein and a high of 2.4 grams of protein, tightly wrapped around the average (which it is not) then you would have concluded (from the *lack* of variation) that there was no need to pursue the data — it doesn't matter: If the variance were this small then the content of your cereal would not depend on your choice of cereal and the protein content of your breakfast cereal would depend more on the size of the bowl than on the choice of the cereal. By contrast, in fact, you found that the variation of protein content was large, with protein content ranging from a low of 0.8 grams of protein to a high of 10.2 grams of protein — which led us forward in search of an explanation.

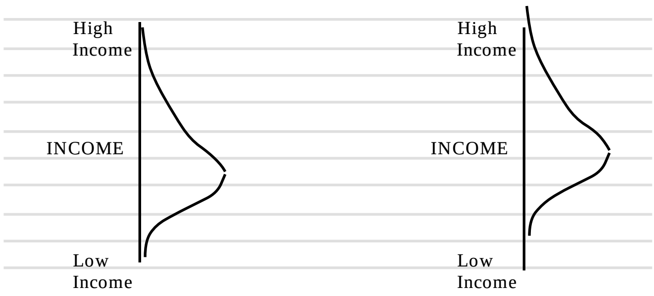
If I were comparing the personal incomes of two groups of people, then again I would have to pay attention to the variation. For example suppose, without real data, that the average income of college graduates from private universities exceeded the average income of graduates from public universities. Even assuming that that is a fact, my evaluation of that fact would depend on the variation: If the variation of incomes in each of the two groups is small, corresponding to the two drawings in Figure 1a, then the difference between public schools and private schools is worthy of examination: Because the variation is small, most of the people in the second group have higher incomes than most of the people in the first group, which would force you to conclude that the difference between the two groups is to be taken seriously. By contrast, if the variation of incomes in each of the two groups is large, corresponding to the two drawings in Figure 1b, then you would conclude that the difference between public schools and private schools is real, but small compared to the overall variation of income. In each case, Figure 1a and Figure 1b, the contrast between the two averages is the same, but your evaluation of these averages depends on the variation.



Income Distribution for Graduates of Public Universities

Income Distribution for Graduates of Private Universities

Comparing Hypothetical Income Distributions with *Small* Variance



Income Distribution for Graduates of Public Universities

Income Distribution for Graduates of Private Universities

Comparing Hypothetical Income Distributions with *Large* Variance

To get this information out of the picture and put it into numbers, we think of the center as a range of data tightly or loosely spread around the average. What remains to be said is exactly what range of data we will report as the “center”.

The Median and the Mean as Centers

I would like you to pretend, for the moment, that you had never heard of a mean or a median and were facing, as if for the first time, the problem of coming up with a number to that represents the center of a collection of numbers. How do you do it? Statisticians have created not so much an answer to this question as they have created a strategy capable of producing answers — sometimes different answers for different occasions.

So I ask — what is the best measure for the center of a distribution of measurements. Specifically, very specifically, what do I mean by “best”?

Well, crudely, the best measure of the center should be close to all of the values in the data. And, of course, that leaves me with the problem of defining close: I'll say that the center, c, is close to the data if the difference between each data point and the center at c, "x_i – c", is consistently small. Actually, this is a bad definition, but I'll follow it out because I want you to see the process of inventing (or re-inventing) the average. So, I define the total deviation from the center c, V(c), by summing up these differences between the data and the center and I write the total variation around each possible value of the center, c, as

$$V(c) = \sum_{i=1}^{n} (x_i - c)$$

First definition of deviation as a function of c.
Total Variation

There is one immediate objection to this measure of deviation: Using this measure a data set with many observations (large "n") will always look worse than a data set with a small number of observations — because n is larger and more deviations get added together into this V.

So, I'll do better, this time correcting for different sizes of "n".

$$V(c) = \frac{1}{n} \sum_{i=1}^{n} (x_i - c)$$

Second definition of deviation as a function of c.
n-adjusted Variation

Is this a good definition? I'll test it by example. Suppose I have three numbers x_1=10, x_2=11, and x_3=12. And suppose I choose 10 as the center. "10" is not the center of 10, 11, and 12, but let me suppose that it is and check out its effect on V. Here, if c=10, V(c)=1

		Center, c, Equal 10
i	x	Difference Between x_i and c
1	10	10-10= 0
2	11	11-10= 1
3	12	12-10= 2
		V(c) = 1

Can I find a better center, a center about which there is less variation? Certainly. Choosing 11 as the center the variation around this center, V(c), is zero. Better. That looks promising. Is there any other "center" around which the variation would be even less?

		Center, c, Equal 11
i	x	Difference Between x_i and c
1	10	10-11=-1
2	11	11-11= 0
3	12	12-11= 1
		E(c) = 0

Unfortunately, yes, there is. Suppose I had tried c =12. That is obviously a bad choice, 12 is not the center of these data, but what does the measure of variation around this center have to say? It says "-1" which may be ridiculous but it is certainly smaller than zero. So, this definition of "error", of the error that results from choosing c as the center, produces ridiculous results — implying that 12 is more central than 11. So — out with the definition. I need a better one.

		Center, c, Equal 12
i	x	Difference Between x_i and c
1	10	10 –12= –2
2	11	11 –12= –1
3	12	12 –12= 0
		V(c) = –1

There are several ways of fixing up the definition, using alternative expressions of what it means for data to be "close" to their center. Suppose I fix up what I did above by saying "No, close is a matter of *distance* not *difference*. I should have used the *distance* between x_i and the center, not the difference." That places "10" a distance of one unit away from 11 and it also places "12" a distance of one unit away from 11. Using this new working definition of variation

$$V(c)=\frac{1}{n}\sum_{i=1}^{n}\left|x_i-c\right|$$

Third definition of deviation as a function of Absolute Deviation

<table>
<tr><th></th><th></th><th>c = 10</th><th>c=11</th><th>c=12</th></tr>
<tr><th>i</th><th>x</th><th>Distance Between x_i and c</th><th>Distance Between x_i and c</th><th>Distance Between x_i and c</th></tr>
<tr><td>1</td><td>10</td><td>|10 –10|=0</td><td>|10 –11|=1</td><td>|10 –12|=2</td></tr>
<tr><td>2</td><td>11</td><td>|11 –10|=1</td><td>|11 –11|=0</td><td>|11 –12|=1</td></tr>
<tr><td>3</td><td>12</td><td>|12 –10|=2</td><td>|12 –11|=1</td><td>|12 –12|=0</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td>V(c)=1</td><td>V(c)=2/3</td><td>V(c)=1</td></tr>
</table>

Sure enough, among these three choices, c=10, c=11, and c=12, 11 is the center with respect to which the variation is smallest. Among these three choices, 11 is the best center.

Minimum Absolute Deviation

In the jargon of the trade, I have found the center "in the sense of minimum absolute deviation" (sometimes abbreviated MAD, *M*inimum *A*bsolute *D*eviation). And the result is, I think, exquisite: I have just defined the median. What's beautiful about it is that I haven't said anything about rank ordering the data, or splitting it in half. I haven't said anything about the median itself. I just defined a measure of "goodness of fit", specifically, minimum absolute deviation, and I said "find the number that is close to the data, close in the sense of minimum absolute deviation." That turns out to be the median. It takes a little bit of calculation to prove that, but it is true and it places the median in context: The median is the "best" measure of the center, "best" in the sense of MAD.

Least Squares

That is not the end of it: The same strategy is able to create other results when it is combined with another definition of "close". The most widely used measure fixes up the difference (definition 1 and 2) by using squares to get rid of the negatives.

$$V(c) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - c\right)^2$$

Fourth definition of deviation as a function of
Squared Deviation

		c=10	c=11	c=12
i	x_i	Squared Distance Between x_i and c	Squared Distance Between x_i and c	Squared Distance Between x_i and c
1	10	$(10 - 10)^2=0$	$(10 - 11)^2=1$	$(10 - 12)^2=4$
2	11	$(11 - 10)^2=1$	$(11 - 11)^2=0$	$(11 - 12)^2=1$
3	12	$(12 - 10)^2=4$	$(12 - 11)^2=1$	$(12 - 12)^2=0$
		V(c)=2.5	V(c)= .667	V(c)=2.5

This too implies that 11 is the center for these hypothetical data. In the jargon of the trade I have found the center "in the sense of least squares". And this logic leads to the mean. The mean is the central value of a distribution "in the sense of least squares".

Homework:
In order of increasing difficulty:

1 I have asserted that *if* variation is defined in the sense of least squares *then* the best number for the center of the variation is the mean. Given the general statement that the variation around a center is minimized when that center is the mean, how small is that minimum variation. That is, what is the value of V(c), V as a function of c, when c is equal to the mean?

$$V(c) = \tfrac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$$

To see the answer (to see the reason for the question) more clearly, the better question is 'what is the *mean* value of the variation around c when c is the mean?' (Substitute $\bar{x}$ into the equation for V and solve for $\frac{1}{n}V$. To see the answer more clearly, when c is the best value (when c is equal to the mean) what is the square root of the mean value of V

2 Using calculus, prove that variance in the sense of least squares, V, is minimized when c = $\bar{x}$ (Differentiate V(c) as a function of c. Set the derivative equal to zero. Solve for c.)

3 Using whatever you can improvise, *prove* that minimum absolute deviation is achieved when c equals the median.

Measuring The Variation

What you have seen is an example of the way statisticians have to be explicit: You can't just say what is the "best"? Not "what is the best way to represent these data?" That's not enough. You have to specify in what sense it is the best.

Once that is done, the rest is "easy": The best average for the data, best in the sense of minimum absolute deviation, is the median. The best average for the data, best in the sense of least squares, is the mean. You will see this strategy at work throughout statistics — when you need the "best" estimate of something.

And how large is the variation? We know that the data varies, but how much? The answer, or answers, to that question are already ordained since the definition of the center used a definition of variation.

So, in the sense of minimum absolute deviation, MAD:

> What is the typical variation around the center? We represent it by average variation around the average or, more specifically, by median of deviations above and below the median. These are the quartiles. Again — I haven't defined this procedurally, telling you how to rank order your data and find quartiles (that will follow). I've given you a strategy whose logical consequence is the quartiles.

And, in the sense of least squares:

> What is the typical variation around the center? We represent it by average variation around the average or, more specifically, by mean deviation around the mean. This is defined as the variance. Unfortunately, squared deviation is a little rough on the human intuition, squared grams of protein for example. So we also define the "standard deviation" as the square root of the variance — putting the deviation in terms that the human intuition can handle.

Variance:

$$V(c) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - c\right)^2$$

Standard Deviation:

$$s_x(c) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - c\right)^2}$$

Average:

…5. transf. The distribution of the aggregate inequalities (in quantity, quality, intensity, etc.) of a series of things among all the members of the series, so as to equalize them, and ascertain their common or mean quantity, etc., when so treated; the determination or statement of an arithmetical mean; a medial estimate. …

6. a. The arithmetical mean so obtained; the medium amount, the generally prevailing, or ruling, quantity, rate, or degree; the `common run.'…

Excerpt from Oxford English Dictionary 2 reference ……

Description: Numbers for the Average

What is a typical income? What's the height of a typical person? What do I mean by "typical"? There are two different numbers in common use as the "typical" values: These are the "median" and the "mean," both of which are sometimes referred to as the "average" (or typical) value.[1] We have looked at the strategy that makes the median and the mean important. The definition of these two typical values is

[1] The language for these things is a little imprecise: Some sources will insist that "average" is the generic term, classifying the median and the mean as two different kinds of average. Other sources will use "average" as the specific term (for the mean). And, of course, it does no good to insist on one definition or the other — because you will encounter both usages in reputable sources. I describe this ambiguity to warn you. For myself, and when necessary, I will use average, the longer word, as the generic term and mean, the shorter word as the specific: The generic term average, will include the median, the mean, and in later chapters a host of additional representations for the central value of a set of data. When possible, I will simply dodge this ambiguity by using generic terms such as "typical value" or "central value," and I will attempt to rely on careful language, in context, to make it clear what is being done.

easy, and generally well known, but for completeness let me now define them and use them.

The Median

Assuming that people come with different incomes, with different heights and weights, assuming that anything we may measure offers a certain amount of variety, what is the typical value? The most useful number for this purpose is the median. (The mean is the number that everyone knows as the "average," but we'll get to that later.) The median of a set of numbers is the one in the middle: If you put numbers in rank order, from low to high, then *the median is the number in the middle: There as many values above the median as there are below it.* Specifically, when the number of elements is odd, the median is the number in the middle. When the number of elements is even, the median is the average of the two numbers in the middle.

For example, in the set of three numbers 1, 2, 3, the median is 2, the number in the middle: There is one number less than 2 and there is one number greater than 2. In the set of four numbers 1, 2, 3, 4, the median is 2.5, the average of 2 and 3: In this set of four numbers there are two numbers less than 2.5 and there are two numbers greater than 2.5.

The only trick, and it's a small one, is to find the middle. Assuming that the numbers are in rank order, we count in to a certain "depth", and that's the median (or the two numbers whose average is the median). For a set of three numbers in rank order we count in to a "depth" of 2 — there's the median. For a set of four numbers we count in to a depth of 2.5, and take the average of the second number and the third.

The formula for the depth is simple: Where "n" is the number of things in the set, the arithmetic is to compute the number (n+1)/2. If the result is a whole number, it identifies the median: If the result is a fraction, then it identifies two numbers whose average is the median.

Working it out, with three things

n = 3 implies the arithmetic (n+1)/2 = 2 So, the depth of the median is 2 and, in rank order, the median is the second number.

With four things

n = 4 implies the arithmetic (n+1)/2 = 2.5 So, the depth of the median is 2.5 and, using the rank order, the median is the mean of the second number and the third.

Going back to breakfast, there are 30 breakfast cereals listed in Figure __. So

n = 30 implies (n+1)/2 = 15.5 The depth of the median is 15.5 and therefore, in rank order, the median is the average of the fifteenth number and the sixteenth. The median number of grams of protein is _____ .

Among the four rice cereals

n = 4 implies (n+1)/2 = 2.5 The depth of the median is 2.5. In rank order, the median is the mean of the second number and the third. The median grams of protein among the rice cereals is _____ .

Among the ___ corn cereals

n = ? implies (n+1)/2 = ?? In rank order, the median is the average of the _____ number and the _____. The median grams of protein among the rice cereals is _____ .

Looking back at those cereals and computing the median protein content you see one important property of the median: For the whole group of cereals, the one lone very high value has little effect on the median; the median lies pretty close to the middle as you see it.

Similarly, for the group of corn cereals, the one odd case, has little effect on the median as a typical value for the group.

There are times when it is possible to be overly precise, and such is the case with the median. Let me use a somewhat more labored definition of the median and then note that a precise definition is not always a specific definition. More precisely, the median is a number such that 50% or more of the data are less than or equal to the median while, at the same time, 50% or more of the data are greater than or equal to the median.

That is more precise, but the nasty fact is that it does not necessarily come up with a single number: In the set of four values, 1, 2, 3, and 4, any number between two and three satisfies the definition of the median. Now, in truth, that's reality: Any number between two and three is equally acceptable as the median of these four numbers. However, convention and simple expectation, "What *is* the median?" seem to be comfortable with a single number, so we improvise to pick one. Unfortunately, even here there is a choice of conventions, none of which is absolutely defensible (and none of which is unreasonable) because, in truth, the median is not always a single number. However, to see the kind of thing you will find, in use, consider the question: What is the median size of a household in the United States? From the U. S. Census Bureau, Current Population Survey, here are the numbers for 1991.

{Need two examples: The first one should have the median fall between two values — that's not what I've got here. Here I have the median category interpreted as an interval, requiring interpolation in the interval.}

Size of Household	Number (in millions) 1991	Percentage Distributio n 1991
1 Person	23.6	25
2 People	30.2	32
3 People	16.1	17
4 People	14.6	15
5 People	6.2	7
6 People	2.2	2
7 or more	1.5	2

Cumulative Percentage Distribution (*Low* to *High*)	Cumulative Percentage Distribution (*High* to *Low*)
25	100
57	**75**
74	**43**
89	**26**
96	**11**
98	**4**
100	**2**

Table ___
Households, by Size of Household: 1991

From the *Statistical Abstract of the United States, 1992*, p. 47, No. 56. On the left, data as presented in the source. On the right, work sheet of cumulative percentages from low to high and from high to low.

The numbers on the left indicate the size of the household and the number of millions of households of that size: There were 23.6 million one-person households, which accounted for twenty-five percent of all households. On the right, you see my work sheet the numbers I need in order to talk about the median.

O.K., now: The median is a number such that 50% or more of the data are less than or equal to the median while, at the same time, 50% or more of the data are greater than or equal to the median.

The cumulative percentages show that the size of 57 percent of households is less than or equal to two while, at the same time, the size of 75 percent of households is greater than or equal to two. So the median size of a household is two.

But if you want to worry about "something better" for this median, you can note it takes almost all of the 30.2 million households to add up to fifty percent (counting up from low to high). That has a feel to it, suggesting that the number is "almost" 3. Looking at it the other way, from high to low, it takes only a few of the 30.2 million households to add up to fifty percent. So, I may choose to interpolate a fraction that will give me a median that is greater than 2, and close to 3. Recognizing that we are dealing with fictions here, useful fictions, but fictions nonetheless: Let me act as if household sizes came in intervals, up to 1.5, 1.5 to 2.5, 2.5 to 3.5, 3.5 to 4.5, and so forth — Acting as if there were some continuous tendency toward small or large households that works itself out in the world as a simple integer, 1, 2, 3, or more.

So now, in this useful fiction, twenty-five percent of households have 1.5 or fewer members. And now, to get up to fifty percent, I need to take most of the next interval: I need another twenty-five percent of the population and the interval contains thirty-two percent of the population, so I need almost all of the population in this group (in order to add up to fifty percent).

So, what do I need? I need 25/32 of the people in the this group. So I will go 25/32 of the way from 1.5 to 2.5 — and call that the median:

Interpolated median (work sheet):

"Interval" Below the Median: 0 to 1.5
End of Interval: 1.5
Percent of Population in Interval Below the Median: 25%

Percent of Population Needed to Add Up to 50%: 25%
"Interval" Including the Median: 1.5 to 2.5
Width of Interval: 1
Percent of Population in Interval Including the Median: 32%
Fraction of this Population needed to Add Up to 50%: 25/32

Interpolated Median:
End of lower Interval plus Interpolated Fraction of Width of Next Interval
= 1.5 plus (25/32) times 1
= 1.5 + (.78)*1
= 2.28

For the sake of increasing my own security with this answer, I check from the other direction: Do I get the same answer?

Interpolated median (work sheet) checking:

"Interval" Above the Median: 2.5 to 7 or more
End of Interval: 2.5
Percent of Population in Interval Above the Median: 43%
Percent of Population Needed to Add Up to 50%: 7%

"Interval" Including the Median: 1.5 to 2.5
Width of Interval: 1
Percent of Population in Interval Including the Median: 32%
Fraction of this Population needed to Add Up to 50%: 7/32

Interpolated Median:
End of Higher Interval Minus Interpolated Fraction of Width of Next Interval
= 2.5 minus (7/32) times 1
= 2.5 + (.22)*1
= 2.28 check!

Result: The interpolated median size of household was 2.3 people.

And now, finally, a word of advice: Don't attempt to memorize these formulas — that's the hard way. Instead, think the problem through from first principles, asking yourself what it is that you are trying to accomplish. And then, just to be careful, check by doing it another way: Here I reasoned from low to high, got an answer, and then did it another way to check: Same answer. If there is any doubt about your method, because there are other acceptable methods, then briefly write it out so your reader will know what you've done. You can't just appeal to authority, saying: "Here it is! This is right." You can't do that because there are different authorities you could appeal to. Each one makes a slightly different interpretation of the problem and it's up to you to choose among them. So, you might as well take the responsibility from the beginning: Think clearly. It's up to you.

Exercise:

Return to the data for the Dow Industrials, showing change as a percentage of the earlier price. Use a stem and leaf diagram to place the industrials in rank order by amount of change and then use that rank ordering to determine the median. Report the "n" (how many industrials are there.?) Report the location for the median. And report the median.

As above, apply the same technique to the rates of infant mortality among states of the United States. Write a brief report of your results.

Exercise: Table _ includes a collection of UNESCO data describing nations of the world, already rank ordered on each of the variables: For gross national product, what is the "n"? For population, what is the "n". Noting that these "n's" are unequal, prepare a very short report stating the median gross national product for nations. Begin at the beginning (although that may not be the first line in your report): Do you trust the data? Are these data complete? What's missing? Do you trust your report of the median?

	Country	**Expectation of Life at Birth (years) 1991**	**Infant Mortality Rate 1991 (Number of deaths of children under 1 year of age per 1,000 live births in a calendar year.)**	**Health Expenditures 1990: per capita on basis of GDP purchasing power parities**		Health Expenditures 1990: % of gross Domestic product (preliminary estimates)	Health Expenditures - Public Health as % of gross domestic product	Health Expenditures - Public Health as % of total health expenditures
1	Australia	**77.0**	**7.9**	**1,151**		7.5	5.2	69.6
2	Austria	**77.3**	**5.4**	**1,192**		8.4	5.6	66.5
3	Belgium	**77.1**	**5.5**	**1,087**		7.4	6.1	82.5
4	Canada	**77.5**	**7.2**	**1,795**		9.0	6.7	74.1
5	Denmark	**75.9**	**6.1**	**963**		6.2	5.2	84.2
6	Finland	**75.8**	**6.0**	**1,156**		7.4	6.2	83.3
7	France	**77.8**	**6.1**	**1,379**		8.9	6.6	74.2
8	Germany	**75.8**	**7.1**	**1,287**		8.1	5.9	72.7
9	Greece	**77.7**	**10.0**	**406**		5.3	4.0	76.0
10	Ireland	**75.5**	**6.2**	**693**		7.1	5.8	82.0
11	Italy	**78.1**	**6.0**	**1,138**		7.7	5.9	75.9
12	Japan	**79.2**	**4.4**	**1,145**		6.5	4.6	71.4
13	Netherlands	**77.8**	**6.9**	**1,182**		8.0	5.8	72.6
14	New Zealand	**75.5**	**9.5**	**853**		7.2	5.9	81.7
15	Norway	**77.1**	**6.7**	**1,281**		7.2	6.9	95.7
16	Portugal	**74.7**	**13.3**	**529**		6.7	4.1	61.7
17	Spain	**78.3**	**6.2**	**730**		6.6	5.2	78.4
18	Sweden	**77.8**	**5.9**	**1,421**		8.7	7.8	89.3
19	Switzerland	**79.1**	**4.7**	**1,436**		7.7	5.2	68.1
20	Turkey	**69.8**	**54.3**	**197**		4.0	1.4	35.6
21	United Kingdom	**76.5**	**7.2**	**932**		6.2	5.2	84.5
22	USA	**75.7**	**10.3**	**2,566**		12.4	5.2	42.4

	Sources:							
		Statistical Abstract of the United States 1991	**Source: Stat Abstact of The United States, 1992, Table 1361, p. 824. "Soviet Union" as former S U "without" Independent Rep**					
		Table 1361 p. 824						
		Original: U.S. Bureau of the Census, World Population Profile: 1991.		**Original: U.S. Bureau of the Census, World Population Profile: 1991.**		Original: U.S. Bureau of the Census, World Population Profile: 1991.	Original: U.S. Bureau of the Census, World Population Profile: 1991.	Original: U.S. Bureau of the Census, World Population Profile: 1991.
				Table 1368 p. 830		Table 1368 p. 829	Table 1368 p. 831	Table 1368 p. 832

				Originl Source Organization of Economic Cooperation and Development, Paris, France, OECD Health Data 1991, and OECD Health Systems: Factgs and Trends, 1993		Originl Source Organization of Economic Cooperation and Development, Paris, France, OECD Health Data 1991, and OECD Health Systems: Factgs and Trends, 1992	Originl Source Organization of Economic Cooperation and Development, Paris, France, OECD Health Data 1991, and OECD Health Systems: Factgs and Trends, 1994	Originl Source Organization of Economic Cooperation and Development, Paris, France, OECD Health Data 1991, and OECD Health Systems: Factgs and Trends, 1995

0	Country	Expectation of Life at Birth (years) 1991
1	Japan	79.2
2	Switzerland	79.1
3	Spain	78.3
4	Italy	78.1
5	France	77.8
6	Netherlands	77.8
7	Sweden	77.8
8	Greece	77.7
9	Canada	77.5
10	Austria	77.3
11	Belgium	77.1
12	Norway	77.1
13	Australia	77.0
14	Israel	77.0
15	Costa Rica	76.8
16	United Kingdom	76.5
17	Denmark	75.9
18	Finland	75.8
19	Germany	75.8
20	USA	75.7
21	Cuba	75.6
22	Ireland	75.5
23	New Zealand	75.5
24	Albania	75.1
25	Singapore	74.8
26	Portugal	74.7
27	Venezuala	74.2
28	Panama	74.0
29	Jamaica	73.6
30	Kuwait	73.6
31	Chile	73.4
32	Yugoslavia	73.0
33	Czechoslovakia	72.9
34	Poland	72.9
35	Bulgaria	72.7
36	Uruguay	72.6
37	Mexico	72.2
38	Romania	71.9
39	Tunisia	71.9
40	Hungary	71.6
41	Jordan	71.2
42	Sri Lanka	71.1
43	Colombia	71.0
44	Argentina	70.9
45	UAR United Arab Emirates	70.9
46	China / People's Republic of China / Mainland	70.0
47	Soviet Union frmr	69.8
48	Turkey	69.8
49	Korea South	69.7
50	Paraguay	69.7
51	Syria	69.4
52	Korea North	69.0
53	Thailand	68.5
54	Lebanon	68.4
55	Libya	68.1
56	Malaysia	68.1
57	Dominican Republic	67.2
58	Iraq	67.0
59	Algeria	66.7
60	Ecuador	66.2
61	Honduras	66.0
62	Saudi Arabia	65.9
63	El Salvador	65.5
64	Brazil	65.2
65	Mongolia	65.1
66	Vietnam	64.7
67	Morocco	64.6
68	Philippines	64.6
69	Iran	64.5
70	Peru	64.3
71	South Africa	64.2
72	Guatemala	63.2
73	Nicaragua	62.5
74	Zimbabwe	61.7
75	Bolivia	61.5
76	Kenya	61.5
77	Indonesia	61.0
78	Egypt	60.8
79	India	57.2
80	Oman	56.6
81	Pakistan	56.6
82	Liberia	56.4
83	Zambia	56.4
84	Somalia	55.9
85	Togo	55.6
86	Papua New Guinea	55.4
87	Senegal	55.1
88	Burma	54.9
89	Ghana	54.6
90	Ivory Coast / Cote d'Ivoire	54.3
91	Congo	54.2
92	Zaire	53.9
93	Haiti	53.6
94	Bangladesh	53.0
95	Sudan	53.0
96	Madagascar	52.6
97	Burkina	52.5
98	Rwanda	52.5
99	Burundi	52.4
100	Tanzania	52.0
101	Ethiopia	51.3
102	Cameroon	51.0
103	Niger	51.0
104	Uganda	51.0
105	Nepal	50.6
106	Benin	50.5
107	Laos	50.2
108	Yemen	49.9
109	Kampuchia / Cambodia	49.3
110	Malawi	49.2
111	Nigeria	48.9
112	Mozambique	47.4
113	Central African Republic	47.1
114	Mali	46.1
115	Sierra Leone	44.8
116	Angola	44.3
117	Afghanistan	43.5
118	Guinea	42.8
119	Chad	39.8
120	Andorra	
121	Antigua and Barbuda	
122	Armenia	
123	Aruba	
124	Bahamas	
125	Bahrain	
126	Barabados	
127	Belize	
128	Bhutan	
129	Bosnia Herzogovina	
130	Botswana	
131	Brunei	
132	Byelarus	
133	Cambodia-cf. Campuchia	
134	Cape Verde	
135	Comoros	
136	Croatia	
137	Cyprus	
138	Djibouti	
139	Dominica	
140	Equatorial Guinea	
141	Estonia	
142	Fiji	
143	Gabon	
144	Gambia	
145	Georgia	
146	Germany East	
147	Germany West	
148	Grenada	
149	Guinea-Bissau	
150	Guyana	
151	Hong Kong	
152	Iceland	
153	Kiribati	
154	Kyrgystan	
155	Lesotho	
156	Liechtenstein	
157	Lithuania	
158	lLatvia	
159	Luxembourg	
160	Maldives	
161	Malta	
162	Mauritania	
163	Mauritius	
164	Moldova	
165	Namibia	
166	Puerto Rico	
167	Qatar	
168	Russia	

169 Saint Kits and Nevis
170 Saint Vincent and the Grenadines
171 San Marino
172 Santa Lucia
173 Sao Tome and Principe
174 Serbia
175 Seychelles
176 Suriname
177 Swaziland
178 Taiwan / Republic of China
179 Tajikistan
180 Trinidad and Tobago
181 Turkmenistan
182 Tuvalu
183 Ukraine
184 Upper Volta
185 USSR
186 Uzbekistan
187 Venuatu
188 Vietnam North
189 Vietnam South
190 Western Somoa
191 Yemen (Aden)
192 Yemen (Sana)

Statistical Abstract of the United States 1991 Table 1361 p. 824 Original: U.S. Bureau of the Census, World Population Profile: 1991.

Mean:

… 8. a. Math. [= F. moyenne, ellipt. for quantité moyenne.] The term (or, in plural, the terms) intermediate between the first and last terms(called the extremes) of a progression of any kind (distinctively, arithmetic(al, geometric(al, harmonic(al mean). Also, in a wider sense, a quantity so related to a set of n quantities that the result of operating with it in a certain manner n times is the same as that of operating similarly with each of the set. In this sense the arithmetic(al mean (commonly called simply the mean) of a set of n quantities is the quotient of their sum divided by n; the geometric(al mean is the nth root of their product. …

(Excerpt from Oxford English Dictionary 2 -- CD -- exact reference??)

The Mean

What is a typical income? What is the size of a typical family? The second number commonly used to answer that question is the mean. Assuming that people come with different incomes, with different heights and weights, the mean is the sum of a set of numbers divided by number of elements in the set.

It may be that the simplicity of the mean and the ease of computation (when you are working without a computer) is what makes it the most commonly used measure of the center: What is the mean of 1 and 2? Add them up, the sum is 3, and divide by 2. The mean of 1 and 2 is 1.5. That’s it. There is nothing more to the arithmetic: Compute the sum of the numbers and then divide the sum into equal parts. Going back to the data reporting protein content of 30 breakfast cereals, listed in Figure __: The sum of the 30 numbers reporting protein content is 89.1 grams of protein. Dividing the sum into 30 equal parts, the average is 89.1 grams of protein divided by 30: The mean is 2.97 grams of protein per serving.

Cereal Number	**Grams of Protein**
1	**4.3**
2	**3.1**
3	**2.9**
4	**2.8**
5	**1.8**
6	**2.2**
7	**3.4**
8	**1.5**
9	**2.1**
10	**5.1**
11	**2.8**
12	**2.7**
13	**10.2**
14	**4.8**
15	**2**
16	**2.6**
17	**3**
18	**2.2**
19	**4.5**
20	**2.1**
21	**1.5**
22	**2.1**
23	**1.6**
24	**0.8**
25	**3.2**
26	**4.4**
27	**2.8**
28	**1.6**
29	**2.2**
30	**2.8**
Sum:	***89.1***
Mean of 30:	***2.97***

You can think of the mean as the answer to a "What if?" question: *What if* the total protein content were divided equally among cereals. *What if* the total income of families were equally divided among families? *What if* the total number of children were divided equally among families? *If* the total were divided into equal parts, the result would be the mean.

Exercises --- as for the median, substituting the mean.

Which Average, The Median or the Mean?

You now have two answers to the question, “What’s typical?, What’s the average?”. You have the median and the mean. And you may suspect, correctly, that there are more answers — which raises the question, “Which one?” When do you use the median? When do you use the mean?

In most cases there is no right answer, no right answer in the sense that you can look up the answer in the back of a book and be correct. There is no clear answer because, usually, the question itself is not clear: “What’s typical? What’s average?” What the person asking the question wants is a general description: How much money do these people have? How large is an average family? Does the average breakfast cereal provide much protein? Is the rate of infant mortality high or low? When you think about what you want from these data, that you want a description, the answer becomes clear: Either number will answer the request for an average that describes the data, as long as you attend to two concerns that will make the description useful: First, be specific about what you’ve done. Say “the median is 2” or “the mean is 3”. (Do not say “the average is 3” or “the average is 2” — that’s not specific.) Second, be consistent: If you or the person reading your work is going to compare the average to some other average, then use the same average. I ask you the average income of families in order to know whether it is higher or lower than it was ten years ago. What average did we use ten years ago? Use the same average now. I ask you the average protein content of rice-based cereals in order to compare rice-based cereals to wheat-based cereals. Use the same average for both.

Sometimes there is an easy answer to that question: What is the average life expectancy of people diagnosed to have a certain disease? Almost certainly, for this case, the correct average is the median because you would have to wait too long to get an answer for the mean: If the disease is detected in 100 people, then when the 51st person has died you will know the median. Happily for your people, but unhappily for your research, you may have to wait much longer to collect all the data you need for the mean.

Sometimes the answer to the question is easy because the median and the average are almost the same. When you find that the stem and leaf diagram gives you a symmetrical result — with the low side of the distribution looking like a mirror image of the high side of the distribution, the median and the average will be almost the same. For example, here is the stem and leaf for infant mortality rates, by state, in the United States. Here, the median, at 9.7 deaths per thousand live births, and the average, at 9.75 deaths per thousand live births, are visually indistinguishable — implying that the difference between the median and the average is small compared to the overall range.

6.0 to 6.499			
6.5 to 6.999	Verm		
7	Haw		
7.5	Minn, Main, Mass		
8	Kans, Utah, RhIsl, NewHam,Wisc, Nev		
8.5	Oregn, Calif, Iowa, Mont, Idah, Conn, Wyo		
9	Neb, WVirg, Okla, Tex, Wash	—	—Median 9.7
9.5	Colo, Ohio, Ariz, NJersy, Penn	<==/—	—Average 9.75
10	NMex, Miss, SDak, Virg		
10.5	NDak, Flor, Kent, Ark, NYork, Tenn		
11	Indian, Louis, Mich		
11.5	Alask, Dela		
12	Alab, SCaro, Miss		
12.5	NCaro, Georg		

Figure _
Stem and Leaf of 1988 Infant Mortality Rates, Showing the Median and the Average.

Numbers shown are in deaths per 1,000 Live Births, see Figure __ of Chapter __.

(Draw the stuff on the right as two arrows pointing from separate labels into one stem.)

But most of the time the best answer is to begin at the beginning: At this stage of your data analysis you are creating a description, so begin with a picture of the thing that is going to be described by the average: Show the stem and leaf (or the histogram, section __). Then compute both averages, both the median and the mean, because the difference between the two, when they are different, improves a better description of the data, better than either one alone.

The result for your "rules" is that I have changed the initial question. The question was "should I use the median or the mean". I've answered by saying, "use both" — because it serves your real purpose, which is to describe the data. And now the new question is "what do I learn by comparing the median to the mean?"

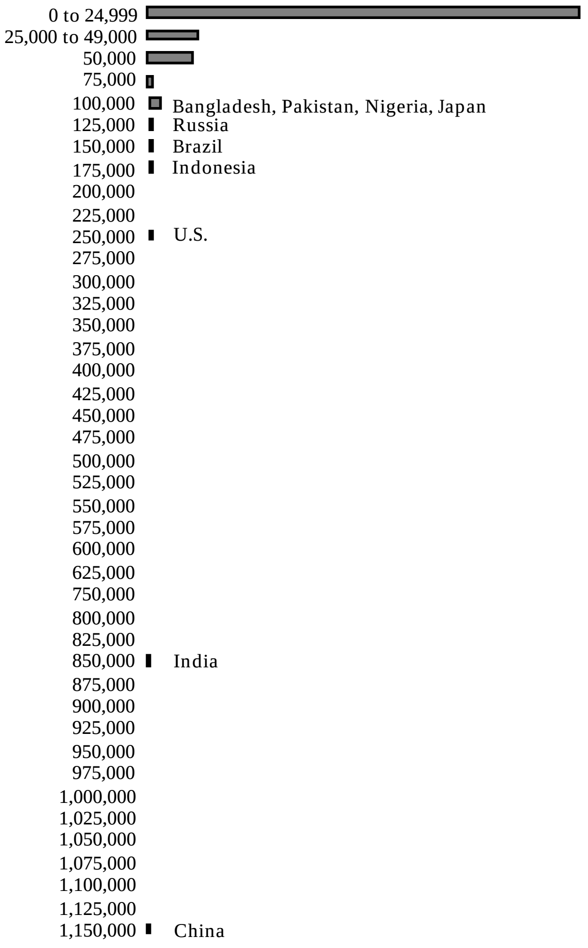
The median is, by definition, the middle value. Therefore, one special feature of the median is that it "ignores" the actual numbers that are above it or below it, as long as they preserve the order so that the same number is in the middle. Thus, for the thirty, breakfast cereals, the median is 2.75 grams of protein, the number in the middle. And if, by chance, by error, by whatever accident, the number for Gerber's had shown up as 5.2 instead of 10.2? The number in the middle is still 2.75. And if the number for Gerber's had been 20.2 instead of 5.2? The number in the middle is still 2.75. And if the number for Gerber's had been 100 times larger than it already is? The number in the middle is still 2.75. That is one property of the median: The median is "robust", with respect to changes in the extreme values.

Rank (Low to High)	Cereal Number (Referring to original alphabetical order)	**Grams of Protein (Rank Ordered)**	**Grams of Protein (Altering the Extreme Value)**	**Grams of Protein (Altering the Extreme Value)**	**Grams of Protein (Altering the Extreme Value)**
1	24	**0.8**			
2	8	**1.5**			
3	21	**1.5**			
4	23	**1.6**			
5	28	**1.6**			
6	5	**1.8**			
7	15	**2.0**			
8	9	**2.1**			
9	20	**2.1**			
10	22	**2.1**			
11	6	**2.2**			
12	18	**2.2**			
13	29	**2.2**			
14	16	**2.6**			
15	12	**2.7**			
16	4	**2.8**			
17	11	**2.8**			
18	27	**2.8**			
19	30	**2.8**			
20	3	**2.9**			
21	17	**3.0**			
22	2	**3.1**			
23	25	**3.2**			
24	7	**3.4**			
25	1	**4.3**			
26	26	**4.4**			
27	19	**4.5**			
28	14	**4.8**			
29	10	**5.1**			
30	13	**10.2**	**5.1**	**20.4**	**1,020.0**
@ rank 15.5	***Median of 30***	***2.75***	***2.75***	***2.75***	***2.75***
	Mean of 30	***2.97***	***2.80***	***3.31***	***36.63***

And thus, when the median is *different* from the mean that's a clue: The median "ignores" the extreme value; the mean uses all values. When they are different, look for extreme values on one side (either above or below) the average. In the case of the breakfast cereals the fact that the median is different from the mean, 2.75 versus

2.97, is a numerical match to what we saw in the stem and leaf — that there is one extreme value, much larger than all the others.

Showing exactly the same properties from a different perspective, consider the data for the gross national products of 150? states.



Rank by Population (Low to High)	Country	Total Population 1990 in thousands (From the Statistica Abstract of the United States, 1991 Table 1359)
1	Tuvalu	9
2	San Marino	23
3	Liechtenstein	28
4	Saint Kits and Nevis	40
5	Andorra	52
6	Antigua and Barbuda	64
7	Aruba	64
8	Seychelles	68
9	Kiribati	70
10	Grenada	84
11	Dominica	85
12	Saint Vincent and the Grenadines	113
13	Sao Tome and Principe	125
14	Santa Lucia	150
15	Venuatu	165
16	Western Somoa	186
17	Maldives	218
18	Belize	220
19	Bahamas	249
20	Barabados	254
21	Iceland	257
22	Djibouti	337
23	Malta	353
24	Equatorial Guinea	369
25	Brunei	372
26	Cape Verde	375
27	Luxembourg	384
28	Suriname	397
29	Comoros	460
30	Qatar	491
31	Bahrain	520
32	Cyprus	702
33	Fiji	738
34	Guyana	753
35	Swaziland	837
36	Gambia	848
37	Guinea-Bissau	999
38	Gabon	1,068
39	Mauritius	1,072
40	Botswana	1,224
41	Trinidad and Tobago	1,271
42	Namibia	1,453
43	Oman	1,481
44	Bhutan	1,566
45	Estonia	1,584
46	Lesotho	1,755
47	Mauritania	1,935
48	Kuwait	2,124
49	Mongolia	2,187
50	Congo	2,242
51	UAR United Arab Emirates	2,254
52	Panama	2,425
53	Jamaica	2,469
54	Liberia	2,640
55	lLatvia	2,695
56	Singapore	2,721
57	Central African Republic	2,877
58	Costa Rica	3,033
59	Uruguay	3,102
60	Albania	3,273
61	Jordan	3,273
62	New Zealand	3,296
63	Lebanon	3,339
64	Armenia	3,357
65	Ireland	3,500
66	Nicaragua	3,602
67	Turkmenistan	3,658
68	Togo	3,674
69	Lithuania	3,726
70	Papua New Guinea	3,823
71	Laos	4,024
72	Sierra Leone	4,166
73	Libya	4,223
74	Norway	4,253
75	Moldova	4,393
76	Kyrgystan	4,394
77	Israel	4,436
78	Bosnia Herzogovina	4,517
79	Paraguay	4,660
80	Benin	4,674
81	Croatia	4,686
82	Honduras	4,804
83	Finland	4,977
84	Chad	5,017
85	Denmark	5,131
86	El Salvador	5,310
87	Tajikistan	5,342
88	Georgia	5,479
89	Burundi	5,646
90	Haiti	6,142
91	Somalia	6,654

92	Switzerland	6,742
93	Bolivia	6,989
94	Kampuchia / Cambodia	6,991
95	Dominican Republic	7,241
96	Guinea	7,269
97	Rwanda	7,609
98	Austria	7,644
99	Senegal	7,714
100	Niger	7,879
101	Tunisia	8,104
102	Mali	8,142
103	Zambia	8,154
104	Angola	8,449
105	Sweden	8,526
106	Bulgaria	8,934
107	Guatemala	9,038
108	Burkina	9,078
109	Malawi	9,197
110	Yemen	9,746
111	Serbia	9,883
112	Belgium	9,909
113	Greece	10,028
114	Byelarus	10,257
115	Portugal	10,354
116	Zimbabwe	10,394
117	Ecuador	10,507
118	Hungary	10,569
119	Cuba	10,620
120	Cameroon	11,092
121	Madagascar	11,801
122	Ivory Coast / Cote d'Ivoire	12,478
123	Syria	12,483
124	Chile	13,083
125	Mozambique	14,539
126	Netherlands	14,936
127	Ghana	15,130
128	Afghanistan	15,564
129	Czechoslovakia	15,683
130	Australia	17,037
131	Saudi Arabia	17,116
132	Sri Lanka	17,198
133	Malaysia	17,556
134	Uganda	18,016
135	Iraq	18,782
136	Nepal	19,146
137	Venezuala	19,698
138	Taiwan / Republic of China	20,435
139	Uzbekistan	20,569
140	Korea North	21,412
141	Peru	21,906
142	Romania	23,273
143	Kenya	24,342
144	Algeria	25,337
145	Morocco	25,630
146	Tanzania	25,971
147	Sudan	26,245
148	Canada	26,538
149	ARgentina	32,291
150	Colombia	33,076
151	Zaire	36,613
152	Poland	37,777
153	SPAN	39,269
154	South Africa	39,539
155	Burma	41,277
156	Korea South	42,792
157	Ethiopia	51,407
158	Ukraine	51,711
159	Egypt	53,212
160	Thailand	56,002
161	France	56,358
162	Iran	57,003
163	Turkey	57,285
164	United Kingdom	57,366
165	Italy	57,664
166	Germany West	63,232
167	Philippines	64,404
168	Vietnam	66,171
169	Germany	79,123
170	Mexico	88,010
171	Bangladesh	113,930
172	Pakistan	114,649
173	Nigeria	118,819
174	Japan	123,567
175	Russia	148,254
176	Brazil	152,505
177	Indonesia	190,136
178	USA	250,410
179	India	852,667
180	China / People's Republic of China / Mainland	1,133,683
	Median	6,398
	Mean	29,707

For these data the median country has 6 million people (6,398,000), like Haiti and Somalia. The mean is 29,707,000, like Argentina. The difference between this median of 6 million and this mean of 30 million is very large. The descriptive solution is to provide both, noting that the difference between the two corresponds to the picture in which we see than there is a "tail" at one end of the distribution. Technically, this is called a "skewed" distribution. And the difference between the median and the mean corresponds to the visual evidence that the population of the world is concentrated in a small number of heavily populated countries

---

Exercise:

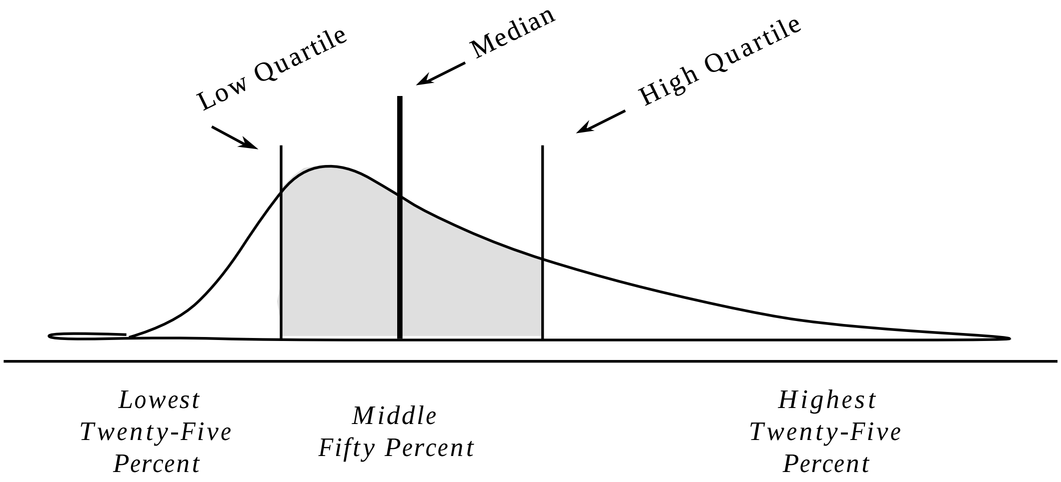
--- Review of previous data, using the stem and leaf, as well as both the median and the mean. Note the differences between the median and the mean. Note the relation between the difference (or the absence of a difference) and the shape of the stem and leaf. Write it up, clearly.

Description: Numbers for the Variation

The Median and The Quartiles

Just as there are several ways of computing the average, there are several ways to compute the variation. But the situation is somewhat simplified because these measures for variation come in pairs. Remember: The median is the center with respect to which variation is minimum in the sense of minimum absolute deviation. So, logically, the median must be paired off with an indicator of absolute deviation. By convention the indicator (or indicators) are the quartiles. (Or, more precisely, by the distances between the medians and the two quartiles.) The two quartiles are average deviations matched to the median: The low quartile is the median of that fifty percent of the data which is below the median. The high quartile is the median of the data that are above the median.) And remember that the mean is the center with respect to which variation is minimum in the sense of least squares. So, logically, if you measure the average as the mean, following the criterion of least squares, then consistency requires that you measure variation by the "variance" which is the mean squared deviation from the mean. (And, in addition, the need to interpret the result in intuition-friendly form will require you to use the standard deviation — which is the square root of the variance -- the square root of the mean squared deviation.) Both these things need to be defined, beginning with the quartiles.

Recall that the median is the middle value. Half of the data are greater than or equal to the median. Half of the data are less than or equal to the median. And now to assess this variation we ask two questions: Among those values that are greater than or equal to the median, what is the average value? And among those values that are less than or equal to the median, what is the average value? And when we have computed those numbers, then the average of the high values and the average of the low values helps us visualize the spread of the data. So we compute the median of those values that are greater than or equal to the median of all values and call it the high quartile. And we compute the median of those values that are less than or equal to the median of all values and call it the low quartile, using the word "*quart*ile" because these three numbers, the low quartile, the median, and the high quartile divide the data into *four* ranges of values. We use these quartiles to visualize the central "hump" of the data.



Hypothetical Income Distribution Divided Into Quartiles
Showing the Middle Fifty Percent of the Data

The range of values between these two quartiles describes the central range of the data:

> *The median protein content of breakfast cereals is __ grams of protein, with the typical breakfast cereal providing between ___ and ___ grams of protein (specifying the quartiles).*

> *The median personal income of these college graduates is ___ (specifying the median), with typical incomes ranging between ___ and ___ (specifying the quartiles).*

That's what we're after, something to express the "middle" of the data, although typically, in print, you will find this information in abbreviated form, simply naming the values: "The median is __, with quartiles at __ and __." That tight little statement, little more than three numbers, presumes that you know what to do with the numbers when you've got them. And what you do with the numbers is to build a mental picture of the center:

> *The median protein content is 2.3 grams, with quartiles at 2.2 grams and 2.4 grams.*

That message gives me a picture of a distribution wrapped tightly around its central value.

> *The median protein content is 2.3 grams, with quartiles at 1.5 grams and 4 .5 grams.*

This message gives me a picture of a distribution that is spread out, and spread more in one direction (toward the high end) than the other.

Customarily we go one step further, adding two more numbers, five in all, to specify the extremes. Thus, completing the description,

> *The median protein content is 2.3 grams, with typical values lying between the low quartile at 1.5 grams and the high quartile at 4.5 grams. In a few instances the values differ considerably from these typical numbers, ranging as low as 0.8 grams, for rice cereals, and as high as 10.2 grams of protein for Gerbers high protein.*

Computing the Quartiles

That's the idea, the rest is detail, important detail, to make sure that we agree on the computation that specifies these quartiles. I will specify a procedure but the important point is the definition: The median divides the data into two sets, high and low. And then the high quartile is the median of the subset of values that are *greater than or equal to* the median. The low quartile is the median of the subset of values that are *less than or equal to* the median.

So to compute these quartiles, we begin as we did with the median, by putting the data in rank order, low to high. Then where "n" is the number of values, the arithmetic is to compute the number (n+1)/2. If the result is a whole number, it identifies the location of the median. If the result is a fraction, then it identifies two numbers whose average is the median

n = number of values in the data

m = location of median = (n+1)/2

If the result is a whole number then the number of values that are greater than or equal to the median is m. And if the result is a fraction, then the number of values that are greater than or equal to the median is the integer part of m, *m*. (If m is 10.5, then its integer part is 10, lopping off the fraction.) And thus the location of the quartile is found by computing the number (m+1)/2. If the result is a whole number, it identifies the location of the quartile. If the result is a fraction, then it identifies two numbers whose average is the quartile.

m = number of values greater than or equal to the median
q = location of quartile = (*m*+1)/2

Exactly the same computation works for the remaining quartile except that you count to q starting at the other end of the distribution. Thus,

m = number of values less than or equal to the median
q = location of quartile = (*m*+1)/2

Working it out with eight things: n = 8 implies the arithmetic (n+1)/2 = 4.5. So, the depth of the median is 4.5 and, using the rank order, the median is the mean of the fourth number and the fifth. The integer part of 4.5 is 4, telling me that the number of values less than or equal to the median is 4.

That gives me *m* = 4. And *m* = 4 implies the arithmetic (*m*+1)/2 = 2.5. So, the depth of the quartile is 2.5 and, using the rank order, the high quartile is the mean of the second and third largest values (in order from large to small) while the low quartile is the mean of the second and third smallest values (in order from small to large).

n = 8
m = 4.5
m = 4
q = 2.5

Working it out with nine things: n = 9 implies the arithmetic (n+1)/2 = 5. So, the depth of the median is 5 and, using the rank order, the median value is the fifth value. The number of values less than or equal to the median is 5.

That gives me *m* = 5. And m = 5 implies the arithmetic (*m*+1)/2 = 3. So, the depth of the quartile is 3 and, using the rank order, the high quartile is the third largest values while the low quartile is the third smallest value (in order from small to large).

n = 9
m = 5
m = 5
q = 3

Working it out with ten things: n = 10 implies the arithmetic (n+1)/2 = 5.5. So, the depth of the median is 5.5 and, using the rank order, the median is the mean of the fifth number and the sixth. The integer part of 5.5 is 5, telling me that the number of values less than or equal to the median is 5.

That gives me *m* = 5. And *m* = 5 implies the arithmetic (*m*+1)/2 = 3. So, the depth of the quartile is 3 and, using the rank order, the high quartile is the third largest value (in order from large to small) while the low quartile is the third smallest value (in order from small to large).

n =10
m = 5.5
m = 5
q = 3

The Mean and the Standard Deviation

The second way to compute variation is paired with the mean. If you measure the average as the mean, then you measure the variation by computing the standard deviation. The idea for the *standard* deviation begins by defining deviation, any deviation, as the difference between a value found in the data and the mean of all the values found in the data. If I have an income of $60,000 and the average income is $50,000, then my deviation is $10,000.

Variance = Mean Squared Deviation $= \frac{1}{n}\sum_{i=1}\left(x_i - \bar{x}\right)^2$ Deviation = Observed Value – Mean Value

Then the basic idea for the *standard* deviation is to compute the mean of the deviations — except that the basic idea doesn't work out. The trouble is that the simple mean of the deviations is a useless number, in fact it is always zero. You can work out this result by simply adding up all the deviations algebraically and dividing by n, computing their mean:

$$\text{Mean Deviation } = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)$$

Following the algebra in steps: Distributing the summation expands the expression for the average to

$$\text{Mean Deviation } = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\bar{x}\right)$$

Evaluating the two expressions on the right, the first is x-bar itself, the mean

$$\text{Mean Deviation} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = \bar{x} - \frac{1}{n}\sum_{i=1}^{n}\left(\bar{x}\right)$$

Evaluating the second expression on the right shows that it too is equal to the mean

$$\text{Mean Deviation} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = \bar{x} - \frac{1}{n}n\bar{x}$$

$$\text{Mean Deviation} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = \bar{x} - \bar{x}$$

which reduces to zero

$$\text{Mean Deviation} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = 0$$

which shows that the average deviation is zero, always zero — so it doesn't tell us anything useful about the data.

The conventional solution is to keep the idea, we are still looking for some sort of average deviation, but modify it by squaring the deviation, computing the mean squared deviation, known as the "variance."

$$\text{Variance} = \text{Mean Squared Deviation} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

That is the basic answer, for a measure of the variation, but we're not quite done with it. There is one more problem: As soon as you try to compute a variance, and then interpret it, you will find that it measures things in squared units: If the data are measured in grams of protein, then the mean squared deviation will give you a result in *square* grams of protein. That's not usable. I can't say "the variance of the protein content of breakfast cereals is 3 square grams of protein." It makes no sense. So, what we do is take the square root and apply the name "standard deviation" written as s_X. Describing the standard deviation in the jargon of the trade, we use the "root mean squared deviation" as the measure of variation with respect to the mean. You can see each of the terms, the root, the mean, and the square, at work in the formula:

$$\text{Standard Deviation of X} = s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Computing the Quartiles

In practice, the way you compute this thing is in stages. Showing the steps for the data on protein content of wheat cereals, Figure __, the first step is to compute the mean: For these five cereals, the sum is 13.8 grams of protein which, when divided into five equal parts gives the mean of 2.76 grams of protein.

Then computing the deviations from the mean, the first datum, 1.6, deviates from the mean by -1.16, for a squared deviation of 1.35. The sum of these squared deviations is 4.35 grams of protein. The variance (the mean squared deviation) is .87 squared grams of protein. And the standard deviation (the root mean squared deviation) is .93 grams of protein.

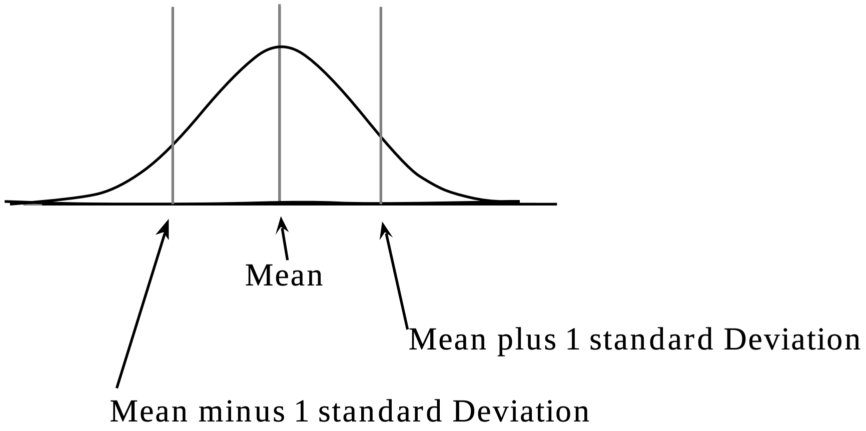
Wheat Cereals	**Protein in Grams**		**Deviations**	**Squared Deviations**
Quaker Puffed Wheat	1.6		-1.16	1.35
Shredded Wheat	2.2		-0.56	0.31
Wheaties	2.8		0.04	0.00
Wheat Chex	2.8		0.04	0.00
Wheat Flakes	4.4		1.64	2.69
Sum	*13.8*			*4.35*
Mean/Variance	***2.76***			***0.87***
Standard Deviation				***0.93***

The rendering of this information into English is traditionally a little opaque:

> *The mean protein content of wheat breakfast cereals is 2.76 grams of protein with a standard deviation of .93 grams.*

> *The mean personal income of these college graduates is ___ specifying the mean, with a standard deviation of __ (specifying the standard deviation.)*

What you are supposed to "see" in that statement is a schematic version of what you actually have. You are supposed to see a distribution of data that is symmetrical and bell shaped: The mean marks the center point. The standard deviation marks off a central region which, schematically, corresponds to the inflection points in the curve of the bell:


Mean
Mean plus 1 standard Deviation
Mean minus 1 standard Deviation

Typically, in writing, one standard deviation is used as a yardstick to mark off small variations while two standard deviations are used to mark off large variations: If the difference between the mean incomes of two different populations is less than one standard deviation, that is taken to suggest that the difference between the means is small (which is not to say that it is unimportant). Two standard deviations are used as a yardstick to mark off large variations: If the difference between the mean incomes of two different populations is more than two standard deviations, that is taken to suggest that the difference between the means is large.

But the real cue to writing and using these things is to keep it simple: You are using the mean and the standard deviation to describe a picture of the data. So, provide the picture, your stem and leaf drawing or a histogram, and accompany it with the numbers. Use them all: Use the median, the quartiles, the mean, and the standard deviation. You will learn, with experience, to match the numbers to the picture, matching the numbers to the peculiar things that are likely to show up in real data. But there is no need to speak in code: Speak, and write clearly. Show the picture. Add the numbers.

Well-Behaved Variables

The Unit of Measure

The Stem and Leaf procedure and its cousin the histogram are powerful devices for making sense out of data: You look at the picture, you construct mental hypotheses by examining first the shape and then the positions of the different units — soy bean cereal is high protein, rice cereal is low protein. Much of the numerical technology invented for the purpose of data analysis provides ways to summarize what you see in these pictures, ways to test the hypotheses you construct, and ways of adding precision to what you have inferred from these pictures. Yet the picture is primary and that is why such a technique as simple as the Stem and Leaf, with little technological challenge, is nevertheless an extremely powerful device in the hands of a data analyst.

However, even here there is a challenge. There are choices to be made. You, the analyst, can not just take the data as given, build a few pictures, report a few numbers, and be done with it — claiming you have analyzed the data. Would that it were that easy.

There are too many pictures that you *could* draw, too many numbers that you *could* compute. So the analyst has to acquire skills that go beyond the routine pictures and mechanical computation of numbers. Which pictures? Which numbers? Why these? And what are their implications. For example, suppose I had set up the breakfast cereal data in terms of grams of protein *per calorie* as the unit of measure instead of setting it up in terms of protein alone.. Using protein *per calorie*, the idea would be to look for high protein qualified by the number of calories that have to be consumed to obtain the measure of protein. Using the ratio of protein to calories would affect the analysis by building-in some compensation for sugar-filled foods as well as air-filled foods that provide neither protein nor anything else in each air-filled serving.

If the idea is to protein to calories, the idea allows at least two implementations. There is a choice: I could compare protein to calories by computing grams of *protein* per *calorie*. Or, I could compare one to the other by computing *calories* per gram of *protein*.

On the face of it, it should make no difference: Whether it is protein per calorie or calories per protein it is the same information — or is it?

1	Rice Krisp Kell; Rice Cerl Gerb;Rice puffed Quak;Rice Flak
2	KixGM; PostToast; BranRaisPost; BranRaisKel; KrumbKell; WheatShred;WheatiesGM; Wheat Chex; MuffetsQuak; BranFlk40%Post; BranFlk40%Kell
3	BranKel; CheerioGM;BarleyGerb;WheatFlkQuak;WheatPuffQuak;Mixed Gerb
4	Oatmeal
5	Special KKell
6	Hi Pro GM
7	
8	
9	
10	HighProGerb

Figure 1a
Grams of Protein per 100 Calories

Stems: 1 gram per 100 calories, 2 grams per 100 calories, ... 10 grams per 100 calories.

10	HighProGerb;HiProGM;SpecialKKell
20	Oatmeal;MixedGerb;WheatPuffQuak;WheatFlkQuak;BarleyGerb
30	Cheerio; BranKel; BranFlk40%Kell; BranFlk40%Post; MuffetsQuak; WheatChex; WheatiesGM; WheatShred; KrumbKell
40	BranRaisKel; BranRaisPost; PostToast; KixGM
50	RiceFlak
60	RicepuffedQuak; RiceCerlGerb; RiceKrispKell

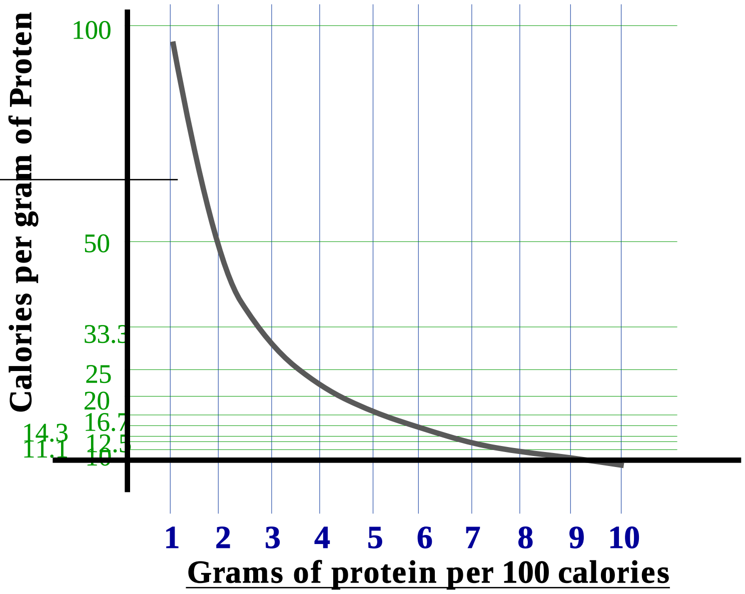
Figure 1b
Calories per gram of Protein
Stems: 10 Calories per gram, 20 calories per gram, ... 60 calories per gram.

The two figures display the two Stem and Leaf Procedures using, in the first case, grams of protein per 100 calories and, in the second case, calories per gram of protein The first S & L, Figure 1a is similar to what we saw earlier with four cases standing out on the high side, one of which, Gerber's High Protein is extreme.

The second S & L is a different shape. There is a dip at 50 followed by a bump at 60, and together these two stems identify four extreme cases. But these are not the same cases — these are all rice cereals. The Gerber's High Protein has migrated to the other end of the distribution. And it no longer stands out, demanding attention. In this picture of the data it is just one leaf on a stem with two other leaves.

This is, for me, "distressing". When I analyzed the breakfast cereal data my intuition and the course of my investigation leaned heavily on what I saw. Here I "see" two different pictures. I have to assume that in the long run whichever picture I use I will end up with the same understanding of the underlying nature of these data — there is one reality behind these data and I had better find it. But it is also clear that the intuitions and the course of my investigation will start off in different directions depending on which picture I use at the beginning of my research.

What's going on? The two different procedures alter the unit of measure. In the first case, one unit, two units, three units, 1, 2, 3, ... counts grams of protein associated with 100 calories. In the second case, one unit, two units, three units counts calories associated with a gram of protein. This has changed the units and, most important, it has changed the intervals between the values.



1 **gram per 100 calories**	↔	100	**calories per gra**
2 grams per 100 calories	↔	50	calories per gra
3 grams per 100 calories	↔	33.3	calories per gra
4 grams per 100 calories	↔	25	calories per gra
5 grams per 100 calories	↔	20	calories per gra
6 grams per 100 calories	↔	16.7	calories per gra
7 grams per 100 calories	↔	14.3	calories per gra
8 grams per 100 calories	↔	12.5	calories per gra
9 grams per 100 calories	↔	11.1	calories per gra
10 grams per 100 calories	↔	10	calories per gra

The change in the unit of measure changes the intervals. That is why Gerber's High Pro has one third of the S&L to itself, in one picture, while Gerber's shares its stem with two other cereals in the other picture of the same data.

In one picture I may say that Gerber's High Pro has approximately double the protein per calorie of its nearest competitor. In the other picture I say that Gerber's High Pro has approximately the same number of calories per gram of protein as General Mills Hi Pro and Kellogg's Special K. Both statements are correct. But they steer your intuition in different directions.

This is the proverbial “tip of the iceberg”. Consider, for example, the data comparing the number of physicians to the number of doctors in a collection of countries. (1975 data).

In physicians per 1,000 people, the median among these 137 countries is 0.385 physicians per thousand people, the number for Mauritania. There is a small number of countries with relatively large numbers of physicians per person among which the USSR and the United Arab Emirates are so extreme as to warrant consideration as special cases. The inner fences establish a “normal” range of variation from -1.5 physicians per 1,000 to 2.8 physicians per thousand.

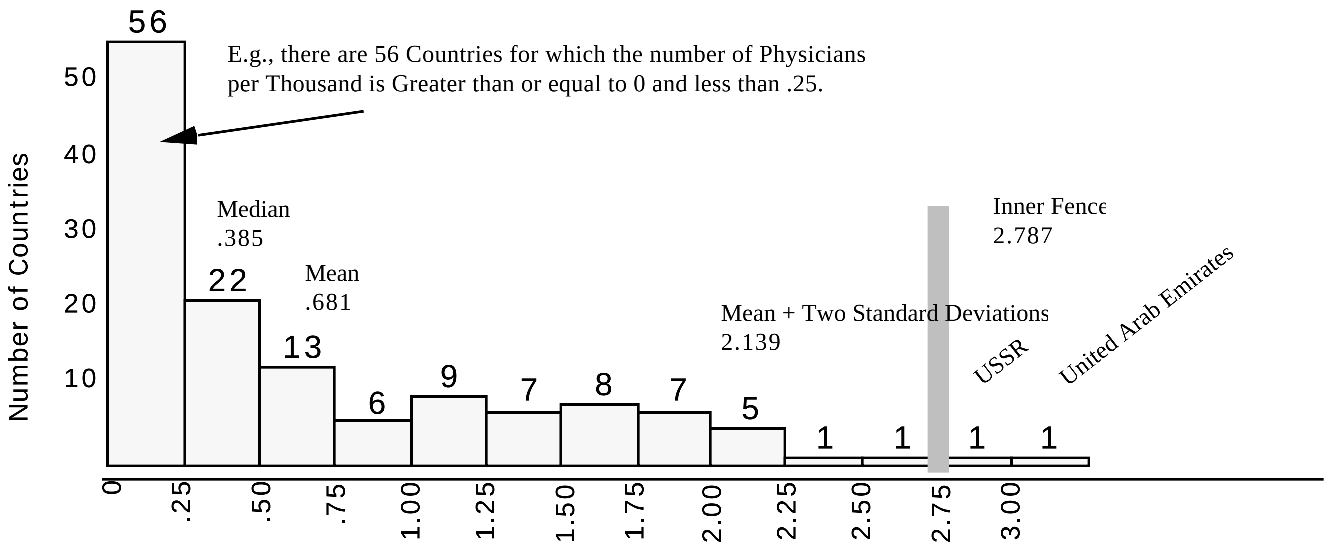


Figure 2

Distribution of 137 Countries with Respect to Physicians per Thousand Population, 1975 Data.

(In physicians per thousand, the median is .385 and the lower and upper quartiles are .077 and 1.161. The mean is .681 and the standard deviation is .729.

That's what the numbers say: They establish a reasonable range from -1.5 to 2.8 physicians per thousand, directing us that nothing within this range is so unusual as to warrant attention as an exception. I have my trouble with any method that directs me not to worry about a negative number of physicians per thousand, that's silly. But ignoring that, they direct my attention to the physician-intensive end for two unusual cases.

That's what the numbers say, or maybe it isn't.

In people per physician, the median among these137 countries is 2,600 people, the number for Mauritania. The inner fences establish a “normal” range of variation from 861 physicians per person to 13,000 physicians per person. The only unusual cases are at the population-intensive end. On the physician-intensive end seven countries exceeding the inner fence and four more countries exceeding the outer fence. Together, the USSR and the United Arab Emirates show the smallest number of people per physician, but neither of these ratios is sufficiently different from adjacent values in the distribution to warrant attention as being different in kind from other countries.

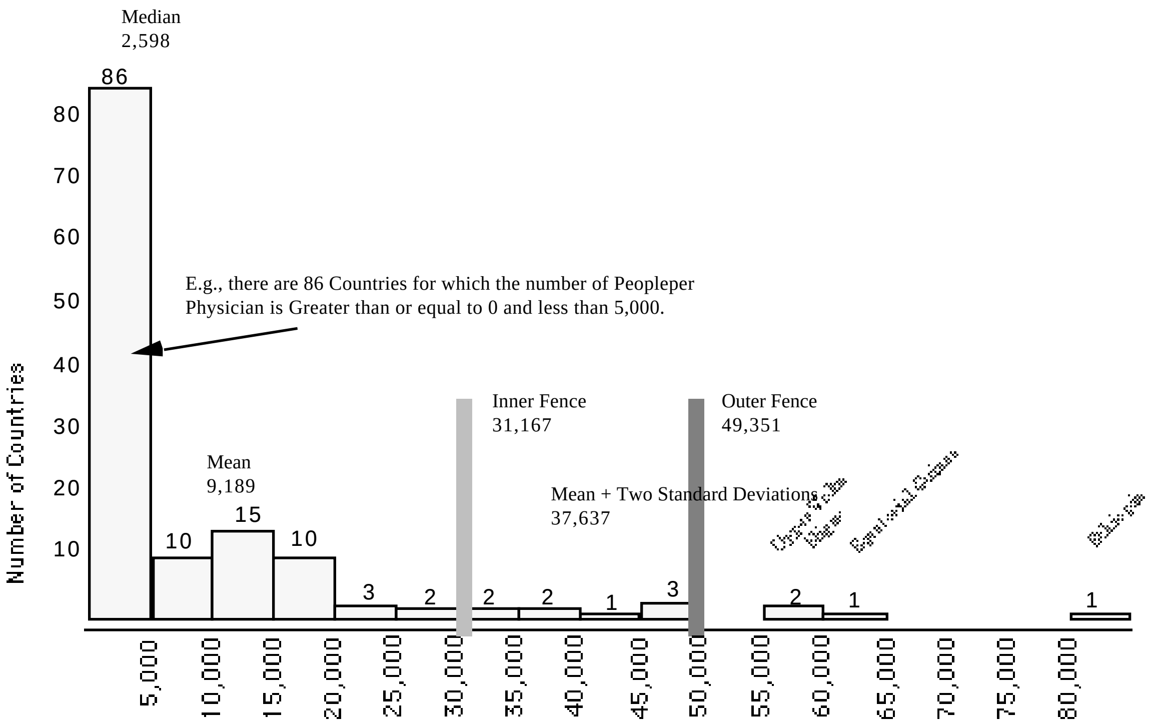


Figure 3

Distribution of 137 Countries with Respect to People per Physician, 1975 Data.

In physicians per person, the median is 2,598 and the lower and upper quartiles are 861 and 12,984. The mean is 9,189 and the standard deviation is 14,224.

This is the kind of now-you-see-it / now-you-don't stuff that gives statistics a bad name — it appears that if I don't like one picture of the data, then I am free to create another. Do I want to use school enrollment data to show that certain schools have excessively large classes — never mind the facts? Then I would use enrollment data organized to count *students per class*. In that form the data will tend to show a "tail" at the high end of the distribution, suggesting that something is wrong or out of line among the largest entries. Do I want to use school enrollment data to show that certain schools have extremely *favorable* faculty student ratios? Then I would organize the same enrollment data to count faculty per student. In that form the data will tend to produce a tail at the opposite end of the distribution where there are large faculty to student ratios.

But this kind of cheating is not data analysis, this is bad data analysis or, perhaps, clever propaganda using numbers and data to create a pretense that its conclusions are objective. This kind of manipulation may be done intentionally, in order to create a picture which is as favorable as possible with respect to the interests of the analyst or the client of the analyst. It may be done unintentionally — because the data were presented in one form and the computations were made on the data as given, without thought for the consequences or the alternatives. Whatever the reason, in the trade we have standards that go a long way toward preventing this kind of lying with statistics, and a long way toward detecting it when it is committed by others. In a phrase, the solution is the "well behaved variable."

	A	B	C	D	E	F	G	H
1	Country	Doctors	Total Population '75		Doctors/Person	Doctors Per 1,000 People		People per Doctor
2	UAR United Arab Emirates	681	220,000		0.003095455	3.095		323
3	USSR	733,744	255,038,000		0.002876999	2.877		348
4	Israel	9,144	3,417,000		0.002676032	2.676		374
5	Czechoslovakia	35,385	14,793,000		0.00239201	2.392		418
6	Bulgaria	18,773	8,793,000		0.002134994	2.135		468
7	Austria	15,702	7,538,000		0.002083046	2.083		480
8	Italy	114,228	55,023,000		0.002076005	2.076		482
9	Greece	18,423	8,930,000		0.002063046	2.063		485
10	Hungary	21,131	10,534,000		0.002005981	2.006		499
11	Germany West	122,069	61,682,000		0.001979005	1.979		505
12	Denmark	9,896	5,026,000		0.001968961	1.969		508
13	Argentina	48,687	25,384,000		0.001918019	1.918		521
14	Belgium	18,510	9,846,000		0.001879951	1.880		532
15	Germany East	31,308	17,127,000		0.001827991	1.828		547
16	Mongolia	2,604	1,446,000		0.00180083	1.801		555
17	Switzerland	11,469	6,535,000		0.001755011	1.755		570
18	Iceland	372	216,000		0.001722222	1.722		581
19	Poland	58,240	33,841,000		0.001720989	1.721		581
20	Norway	6,884	4,007,000		0.001717994	1.718		582
21	Canada	39,104	22,801,000		0.001715012	1.715		583
22	Sweden	14,045	8,291,000		0.001694006	1.694		590
23	USA	348,484	213,925,000		0.001629001	1.629		614
24	Netherlands	21,826	13,599,000		0.001604971	1.605		623
25	SPAN	54,992	35,433,000		0.001552	1.552		644
26	France	77,888	52,913,000		0.001472001	1.472		679
27	Finland	6,699	4,652,000		0.001440026	1.440		694
28	New Zealand	4,110	3,031,000		0.001355988	1.356		737
29	Romania	28,548	21,178,000		0.001348003	1.348		742
30	United Kingdom	75,612	56,427,000		0.001339997	1.340		746
31	Yugoslavia	27,143	21,322,000		0.001273004	1.273		786
32	Portugal	11,101	8,762,000		0.001266948	1.267		789
33	Ireland	3,773	3,131,000		0.001205046	1.205		830
34	Japan	133,344	111,120,000		0.0012	1.200		833
35	Puerto Rico	3,479	2,902,000		0.001198828	1.199		834
36	Malta	382	329,000		0.001161094	1.161		861
37	Libya	2,586	2,255,000		0.001146785	1.147		872
38	Luxembourg	368	342,000		0.001076023	1.076		929
39	Venezuala	13,105	12,213,000		0.001073037	1.073		932
40	Qatar	96	90,000		0.001066667	1.067		938
41	Kuwait	1,089	1,085,000		0.001003687	1.004		996
42	Cuba	8,201	9,481,000		0.000864993	0.865		1,156

	A	B	C	D	E	F	G	H
43	Paraguay	2,229	2,647,000		0.000842085	0.842		1,188
44	Panama	1,404	1,678,000		0.00083671	0.837		1,195
45	Cyprus	547	673,000		0.000812779	0.813		1,230
46	Bahamas	161	200,000		0.000805	0.805		1,242
47	Lebanon	2,301	2,869,000		0.000802022	0.802		1,247
48	Togo	1,623	2,248,000		0.000721975	0.722		1,385
49	Peru	10,514	15,326,000		0.000686024	0.686		1,458
50	Hong Kong	2,881	4,225,000		0.000681893	0.682		1,467
51	Bahrain	177	260,000		0.000680769	0.681		1,469
52	Barabados	166	245,000		0.000677551	0.678		1,476
53	Costa Rica	1,292	1,994,000		0.000647944	0.648		1,543
54	Nicaragua	1,400	2,318,000		0.000603969	0.604		1,656
55	Brazil	62,656	109,730,000		0.000571002	0.571		1,751
56	Trinidad and Tobago	550	1,009,000		0.000545094	0.545		1,835
57	Turkey	21,696	39,882,000		0.000544005	0.544		1,838
58	Mexico	31,556	59,204,000		0.000533005	0.533		1,876
59	Korea South	17,851	34,663,000		0.000514987	0.515		1,942
60	Colombia	12,997	25,890,000		0.000502008	0.502		1,992
61	Ecuador	3,517	7,090,000		0.000496051	0.496		2,016
62	South Africa	12,060	24,663,000		0.000488992	0.489		2,045
63	Suriname	202	422,000		0.000478673	0.479		2,089
64	Bolivia	2,581	5,410,000		0.000477079	0.477		2,096
65	Dominican Republic	2,375	5,118,000		0.000464048	0.464		2,155
66	Vietnam South	9,000	19,650,000		0.000458015	0.458		2,183
67	Chile	4,419	10,253,000		0.000430996	0.431		2,320
68	Iraq	4,504	11,067,000		0.000406976	0.407		2,457
69	Saudi Arabia	3,613	8,966,000		0.000402967	0.403		2,482
70	Mauritius	346	899,000		0.000384872	0.385		2,598
71	Seychelles	21	60,000		0.00035	0.350		2,857
72	Iran	11,358	32,923,000		0.000344987	0.345		2,899
73	Western Somoa	55	160,000		0.00034375	0.344		2,909
74	Syria	2,403	7,259,000		0.000331037	0.331		3,021
75	Philippines	13,464	44,437,000		0.000302991	0.303		3,300
76	Honduras	920	3,037,000		0.000302931	0.303		3,301
77	Guyana	237	791,000		0.000299621	0.300		3,338
78	Jamaica	570	2,029,000		0.000280927	0.281		3,560
79	Jordan	745	2,688,000		0.000277158	0.277		3,608
80	El Salvador	1,117	4,108,000		0.000271908	0.272		3,678
81	Pakistan	17,922	70,560,000		0.000253997	0.254		3,937
82	Grenada	25	100,000		0.00025	0.250		4,000
83	India	145,946	613,217,000		0.000238001	0.238		4,202
84	Sri Lanka	3,245	13,986,000		0.000232018	0.232		4,310
85	Egypt	8,034	37,543,000		0.000213995	0.214		4,673

	A	B	C	D	E	F	G	H
86	Tunisia	1,213	5,747,000		0.000211067	0.211		4,738
87	Oman	153	770,000		0.000198701	0.199		5,033
88	Guatemala	1,207	6,129,000		0.000196933	0.197		5,078
89	Gabon	96	521,000		0.000184261	0.184		5,427
90	Burma	5,561	31,240,000		0.000178009	0.178		5,618
91	Malaysia	2,007	12,093,000		0.000165964	0.166		6,025
92	Congo	213	1,345,000		0.000158364	0.158		6,315
93	Sao Tome and Principe	12	80,000		0.00015	0.150		6,667
94	Zimbabwe	916	6,272,000		0.000146046	0.146		6,847
95	Swaziland	65	469,000		0.000138593	0.139		7,215
96	Thailand	5,009	42,093,000		0.000118998	0.119		8,403
97	Ghana	938	9,873,000		9.50066E-05	0.095		10,526
98	Kenya	1,246	13,251,000		9.40306E-05	0.094		10,635
99	Madagascar	754	8,020,000		9.4015E-05	0.094		10,637
100	Zambia	470	5,004,000		9.39249E-05	0.094		10,647
101	Botswana	63	691,000		9.11722E-05	0.091		10,968
102	Haiti	396	4,552,000		8.69947E-05	0.087		11,495
103	Liberia	142	1,708,000		8.31382E-05	0.083		12,028
104	Sudan	1,407	18,268,000		7.70199E-05	0.077		12,984
105	Maldives	9	120,000		0.000075	0.075		13,333
106	Morocco	1,243	17,504,000		7.10123E-05	0.071		14,082
107	Senegal	305	4,418,000		6.90358E-05	0.069		14,485
108	Bangladesh	5,088	73,746,000		6.89936E-05	0.069		14,494
109	Comoros	21	306,000		6.86275E-05	0.069		14,571
110	Mauritania	87	1,283,000		6.78098E-05	0.068		14,747
111	Nigeria	4,224	63,049,000		6.69955E-05	0.067		14,926
112	Ivory Coast / Cote d'Ivoire	322	4,885,000		6.59161E-05	0.066		15,171
113	Guinea	278	4,416,000		6.29529E-05	0.063		15,885
114	Indonesia	8,299	136,044,000		6.10023E-05	0.061		16,393
115	Somalia	193	3,170,000		6.08833E-05	0.061		16,425
116	Angola	384	6,394,000		6.00563E-05	0.060		16,651
117	Yemen (Sana)	367	6,668,000		5.5039E-05	0.055		18,169
118	Cameroon	354	6,433,000		5.50288E-05	0.055		18,172
119	Tanzania	846	15,388,000		5.49779E-05	0.055		18,189
120	Mozambique	507	9,223,000		5.49713E-05	0.055		18,191
121	Central African Republic	97	1,790,000		5.41899E-05	0.054		18,454
122	Singapore	106	2,248,000		4.7153E-05	0.047		21,208
123	Laos	155	3,303,000		4.6927E-05	0.047		21,310
124	Lesotho	49	1,148,000		4.26829E-05	0.043		23,429
125	Uganda	431	11,353,000		3.79635E-05	0.038		26,341
126	Afghanistan	656	19,280,000		3.40249E-05	0.034		29,390
127	Zaire	807	24,450,000		3.30061E-05	0.033		30,297

	A	B	C	D	E	F	G	H
128	Benin	95	3,074,000		3.09044E-05	0.031		32,358
129	Nepal	339	12,572,000		2.69647E-05	0.027		37,086
130	Rwanda	106	4,233,000		2.50413E-05	0.025		39,934
131	Mali	142	5,697,000		2.49254E-05	0.025		40,120
132	Burundi	83	3,765,000		2.20452E-05	0.022		45,361
133	Chad	83	3,947,000		2.10286E-05	0.021		47,554
134	Malawi	103	4,909,000		2.09819E-05	0.021		47,660
135	Upper Volta	109	6,032,000		1.80703E-05	0.018		55,339
136	Niger	83	4,600,000		1.80435E-05	0.018		55,422
137	Equatorial Guinea	5	313,000		1.59744E-05	0.016		62,600
138	Ethiopia	338	28,134,000		1.20139E-05	0.012		83,237
139								
140					Median	0.385		2,598
141					Low Q	0.077		861
142					High Q	1.161		12,984
143								
144					Spread	1.084		12,122
145					Step Size	1.626		18,184
146								
147					Low inner fence	-1.549		-17,322
148					High inner fence	2.787		31,167
149								
150					Low Outer Fence	-3.175		-35,506
151					High Outer Fence	4.413		49,351
152								
153								
154					Mean	0.681		9,189
155					Standard Dev	0.729		14,224
156					Mean-2sd	-0.777		-19,260
157					Mean-2sd	2.139		37,637

objective ...determined by and emphasizing the features and characteristics of the object, or thing dealt with, rather than the thoughts, feelings, etc., or the artist, writer, or speaker.

subjective ...of or resulting from the feelings or temperament of the subject, or person thinking, rather than the attributes of the object thought of ...

Webster's New World Dictionary, College Edition1956.

Transforming the Complex into the Simple: Well-Behaved Variables

Data analysts attempt to reach objective conclusions. But the path to objective results is full of seeming contradictions one of which is that the path itself is not objective, only the conclusion. If there is one way of assigning a unit of measure to the unit of analysis, then there is *always* a second, and a third and infinitely many way of assigning units of measure. And the analyst must choose among them. The choice will have consequences. It will affect the path of the research. Yet the choice must be made before the result is known. So — among the many ways to proceed with one set of data — which one is right?

The answer depends on the concept of a *well-behaved variable.* Eventually I will provide reasons why this concept "should" be as useful as it is. Eventually, I will philosophize with respect to its meaning. But make no mistake: The "proof of the pudding" is that this thing — the concept of a well-behaved variable — works. Logical argument as to why this concept should work may or may not be convincing. My explanation of the reason why this concept works may or may not even be correct. No matter. It works.

Five properties identify “well-behaved” variables. A well-behaved variable is:

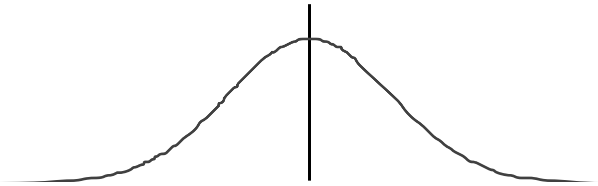
1. Symmetrical
2. Homeoscedastic
3. Linear
4. Additive

and

5. It makes sense.

1. Symmetrical

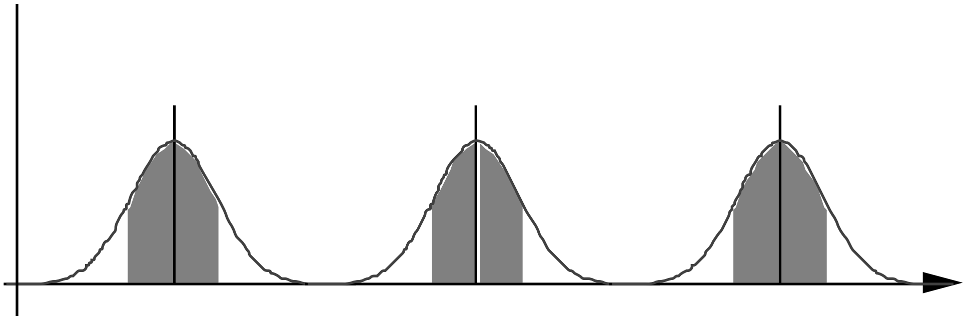
The distribution of a well-behaved variable is symmetrical around the center of the distribution: The upper quartile is as far above the median as the lower quartile is below the median. Often a well behaved variable is both symmetrical and bell-shaped, suggesting an idealized form known by several names as a “bell shaped curved”, or “normal distribution” or “Gaussian distribution”.



2. Homeoscedastic

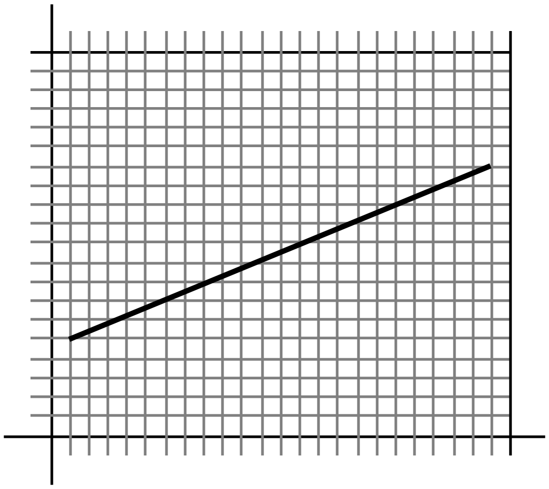
The variation of a well-behaved variable is constant from case to case. For example, if individual wealth is a well-behaved variable, then the variation of wealth in the United States in 1960 and the variation of wealth in the United States in 1990, must be (more or less) constant from 1960 to 1990 — even though the average income will have increased considerably during those thirty years. *If* individual wealth is well-behaved, then although the average income will have changed between 1960 and 1990, the variation will not.

Similarly, if individual wealth is a well-behaved variable, then the variation in wealth among those who have completed high school and the variation in wealth among those who have completed college must be (more or less) the same. The average income of the college graduate will exceed the average income of the high school graduate, but the variation of income within each educational group will be the same.



3. Linear

If two well-behaved variables are related at all, then the relation between two well-behaved variables is likely to be linear — This becomes important later when we look for relations and correlations between variables.



4. Additive

If a variable is well-behaved then effects that serve to increase or decrease its value will decrease it or increase it additively. Perhaps the most commonly used examples of non-additive variables are related to health where it is often suggested that risk factors (e.g., smoking, lack of exercise, overweight, poor diet, heredity as risk factors associated with cardiovascular disease) are spoken of as multiplicative in their consequences for disease. This too becomes important later.

5. Makes Sense

If the *logarithm* of personal income is well-behaved, then the logarithm will have an interpretation and it will make sense.

If the cube root of the weight of organisms is well-behaved, while the weight itself is not, then the cube root of weight is a correct unit of measure. The cube root will have an interpretation and there will be good reason why the cube root makes sense.

Memorize this list: *Symmetrical*, *homeoscedastic*, *linear*, *additive*, and *correct*. Much of the magic an experienced data analysis can perform, much of our ability to go beyond common sense, to find order in data, and then to make sense of it, depends on the use of variables that are *symmetrical*, *homeoscedastic*, *linear*, *additive*, and *correct* . That is to say, much of the power that a data analyst and exercise depends on well-behaved variables .

Exercises

1. Compute means and standard deviations from useful subsets of the data for breakfast cereals. Does protein content appear to be homeoscedastic?

2. Ditto -- the ten gram weight

3. Tukey Viscosity. Tukey, *Exploratory Data Analysis*, page 25, quoting

> In 1963, McGlanery and Harban gave the values in panel A, showing how well they could measure the viscosity of liquids with a device called a capillary rheometer. Make appropriate stem-and-leaf displays for each of the three samples; comment on the appearance of each.

Run (for)	Viscosity in 100,000’s of poises		
each sample	Sample I	Sample II	Sample III
1	.384	.661	3.54
2	.376	.671	3.66
3	.376	.688	3.42
4	.371	.644	4.10
5	.385	.668	4.09
6	.377	.648	3.77
7	.365	.706	4.17
8	.384	.715	3.91
9	.365	.647	4.61
10	.384	.682	3.87
11	.378	.692	
12		.729	

(Original source: R. M. McGlanery and A. A. Harban 1963, “Two instruments for measuring the low-sheer viscosity of polymer melts,” *Materials Research and Standards 3:* 1003-1007. Table 2 on page 1004.)

4. Get the income data referred to in the text.

Transformation

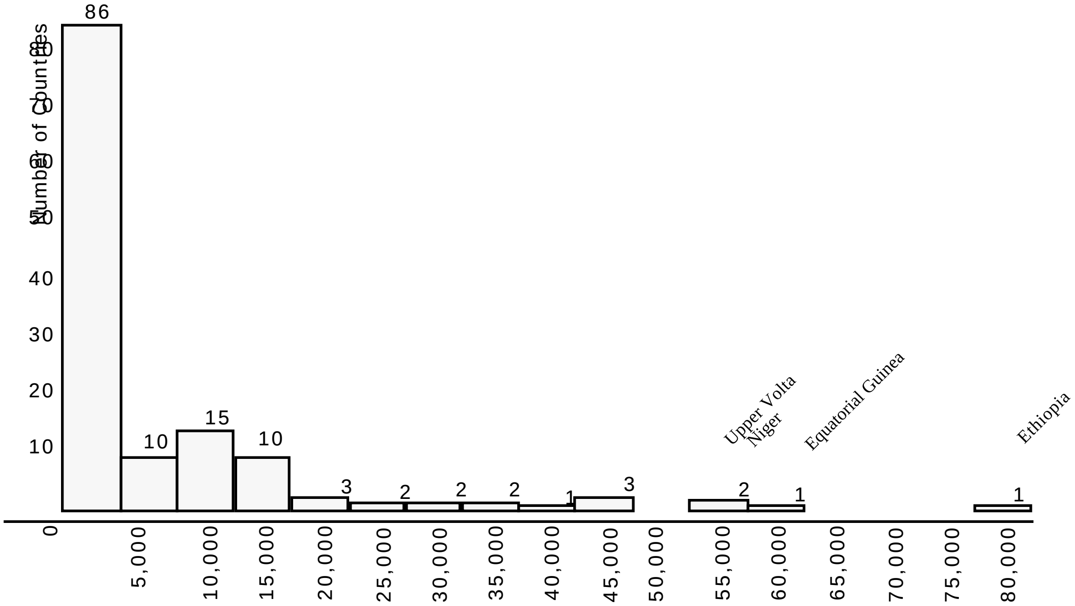
Whenever possible data analysts do not fit their tools to the data, we fit the data to our tools. Data analysts *could* develop tools for variables that are not well behaved, for variables whose distribution is not symmetrical, for relations that are not linear, and so forth, but we do not. Instead we transform the data to make it well behaved.

Mathematicians do something similar and it gives them great power: When you have a relatively "difficult" problem like a multiplication problem, a transformation changes a multiplication problem to an addition problem (applied to the logarithms). When you have a difficult problem like the analysis of the sound wave of a musical instrument, a transformation changes combinations of sine waves and cosine waves into linear combinations (of their Fourier transforms).

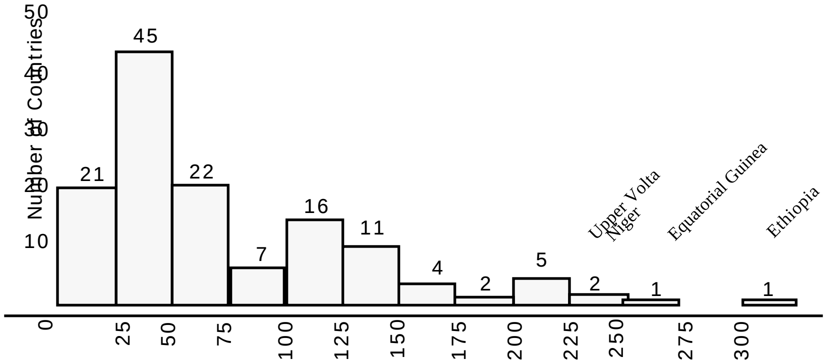
It is a general strategy for handling complicated problems. Instead of tackling them head on, the genius of the mathematics is to figure how to transform the problem into something simple. The rest is (relatively) easy. Data analysts uses the same strategy — transforming the unit of measure in order to create a well-behaved variable. After that the rest of the analysis is easier.

For example, here is a preview of coming attractions: Figure _ shows five transformations of people per physician, transforms ranging from the identity transformation (identical to the variable as given) to the inverse transformation. Between the two extremes you see three intermediate results corresponding to a square root transformation, a logarithmic transformation, and an inverse square root transformation, five in all. (Remember — we don't have to make sense out of all of these things, only the one that is well behaved. I don't have to make sense out of the square root or the inverse square root of people per physician — unless it is well-behaved.)

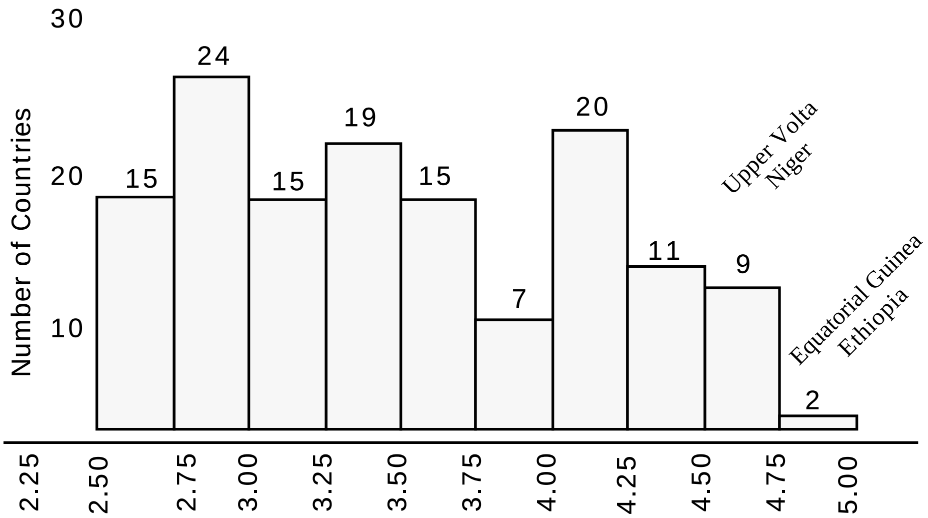
Look at the shapes of these five pictures of the data. In sequence the histograms show a systematic change in the behavior of the distributions. The first, the identity transformation is asymmetrical with four physician poor countries, Ethiopia, Equatorial Guinea, Niger, and Upper Volta in extreme positions, out on the tail, away from the main body of the data (without a corresponding tail at the other end of the distribution.



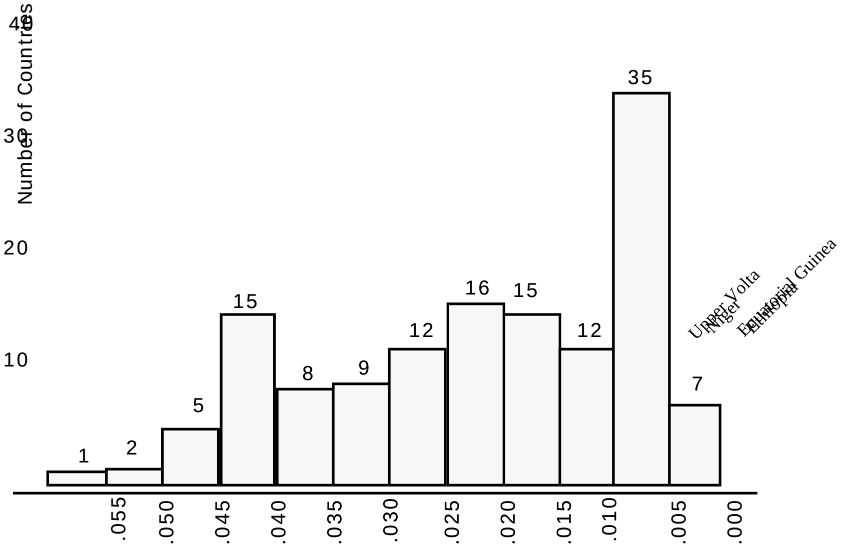
Histogram of Persons per Physician



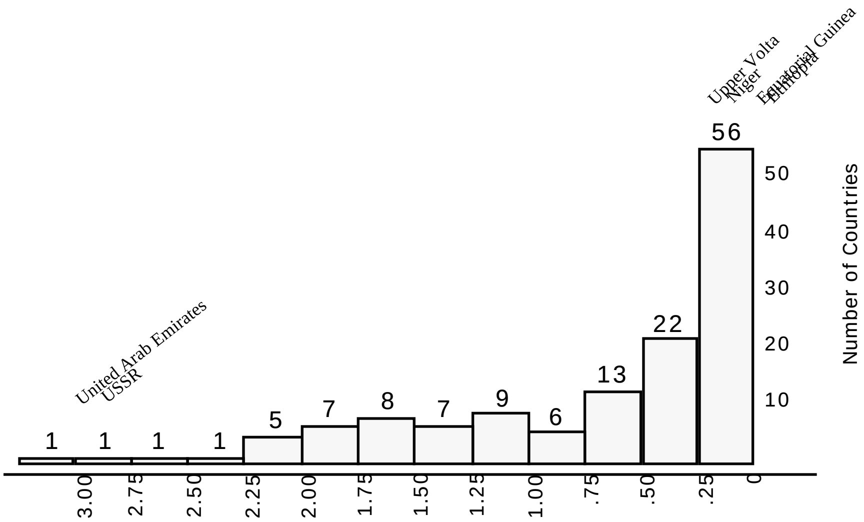
Histogram for Square Root of Persons per Physician



Histogram for Log (base 10) of Persons per Physician



Histogram for Inverse Square Root of Persons per Physician



Histogram for Inverse of Persons per Physician , i.e., for Physicians per Person
(Recorded as Physicians per 1,000 Persons)

The next transformation, using the 1/2 power instead of the first power is less extreme. The behavior has grown a tail on the left while the tail on the right is less extreme.

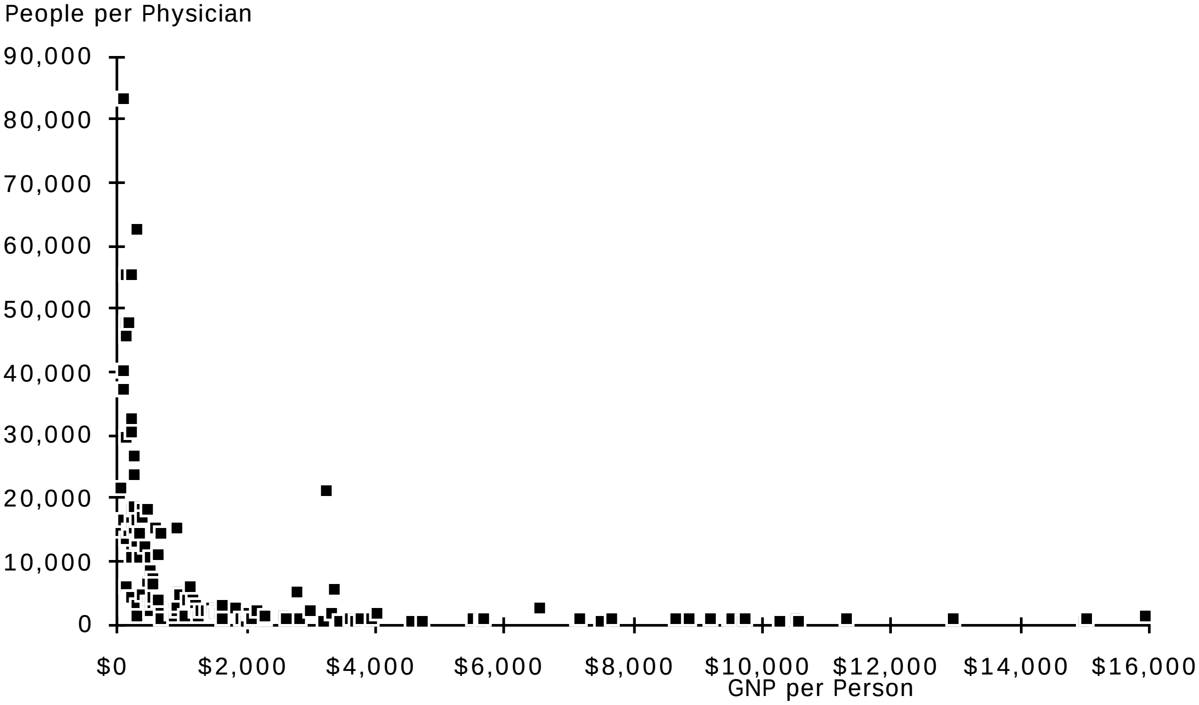
The next transformation using the logarithm instead of the .5 power is (relatively) symmetrical.

The next transformation, using the -.5 power of the variable, shows a long tail on the left and a short tail on the right. For example, the two countries that were previously extreme are now packed in close to the center.

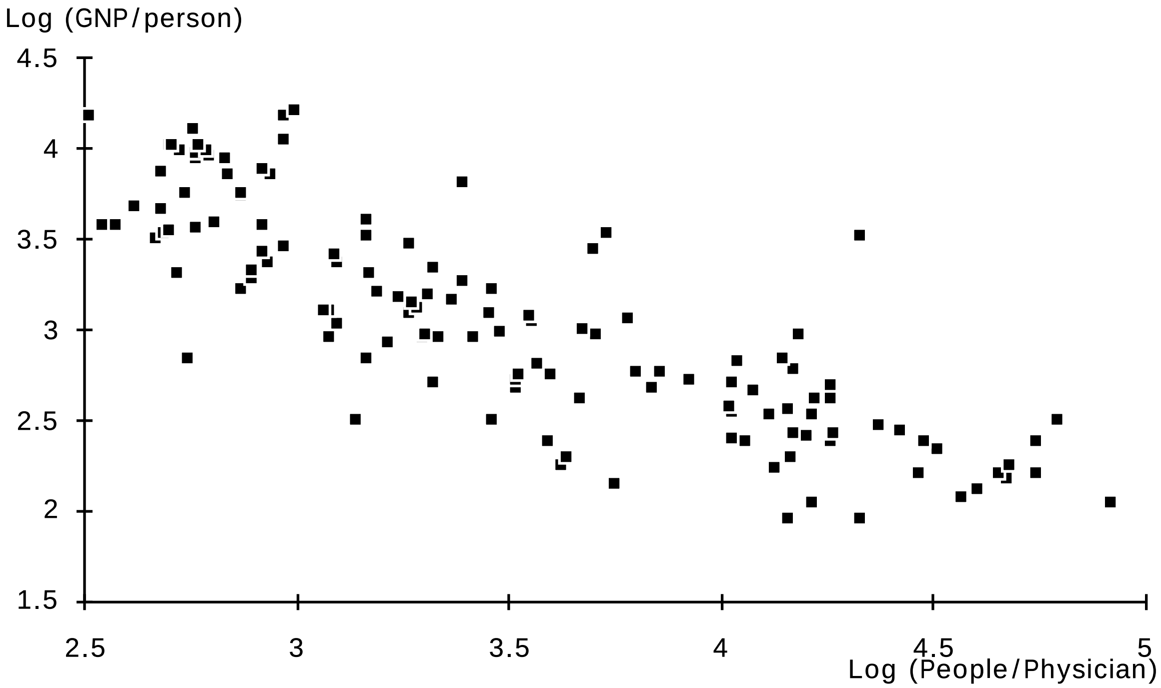
And finally the inverse transformation shows extreme behavior. It is decidedly asymmetrical, but here the tail is tacked to the opposite end as compared to the behavior of the original unit of measure.

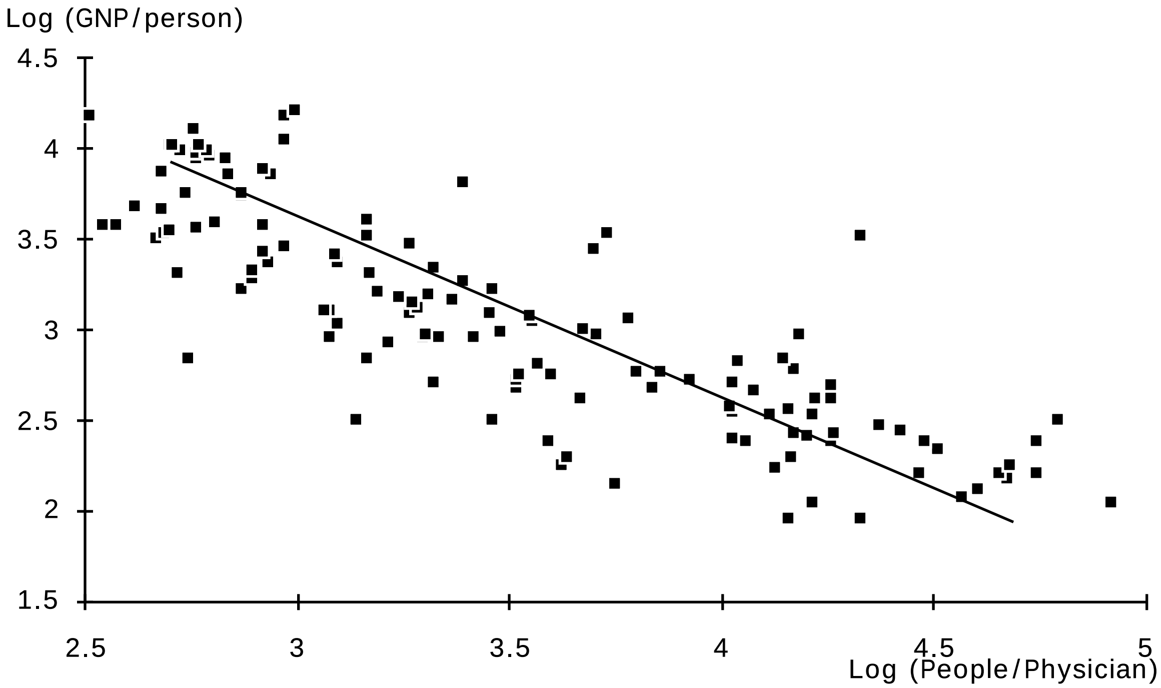
That is a demonstration of the power of transformation to change the picture. From these the data analyst chooses how the data will be and how the data should be transformed.

For another preview, consider physicians per person, the same variable, compared to GNP per capita for the same countries. Figure _ shows the graph of the relation between these two variables. Neither variable is well behaved. Neither is symmetrical (first criterion). And their relation is not well behaved — not linear (third criterion). I would not care to go forward with an analysis of the relation displayed in this graph — too complicated: The picture suggests a chevron shaped distribution with two distinct wings.



Now, transforming both variables, here is the new picture of the relation (using the same data):


Log (GNP/person)
4.5
4
3.5
3
2.5
2
1.5
2.5
3
3.5
4
4.5
5
Log (People/Physician)



This picture gives me a place to begin. Both figures "describe" exactly the same reality, but in the second figure provides my intuition and my sense with ample cues with which to interpret the relation. The second picture is approximately linear. Moreover, the slope is close to being negative one — not some relatively complicated number like 2 or 3, but negative 1. I can begin to make sense of that. (It tells me that to a first approximation the number of physicians in the country is proportional to the wealth of the country.

The most common choices among transformations are organized according to the power of the transformation, where power refers to the exponent of the transformation. Here we have considered five, where the "0" power is considered the log. The five transformations had a progressive effect. An increase of power decrease the appearances of tails on the right and increases the appearances of a tail on the left. A decrease in power increases the appearances of tails on the left and decreases the appearances of a tail on the right.

Power	Transformation		May Indicate
…	…		
1	x^1	Identity transformation (no change)	
1/2	$x^{1/2}$	Square Root	Counts / Waiting times
0	log(x)	Logarithm	Amounts. Wealth, Counts
-1/2	$x^{-1/2}$	Inverse square root	
-1	x^{-1}	Inverse	Rates
…	…		

Common Transformations and Their Indication

For p=1 the picture shows a highly asymmetrical distribution with the United Arab Emirates and USSR out on the tail away from the main body of the data. By contrast, for p=-1 the behavior is highly asymmetrical in the opposite direction, with Ethiopia, Equatorial Guinea, Niger, and Upper Volta solidly close to the main body of the data — the tail is at the other end. In between there is a transition from one extreme shape to the other. And we narrow the choice among alternative units of measure by choosing the one that is well-behaved.

Interpreting The Data

For people per physician, the well-behaved choice requires a compromise between the first criterion, symmetry, and the fifth criterion, sense. By itself, the first criterion would lead to something like the negative one tenth power, p = -.1. But the fifth criterion leaves me in trouble attempting to interpret the -.1 power whereas the logarithm, which is close, is easy to interpret. So I will use the logarithm of physicians per capita as my unit of measure. The sense of the logarithm is that adding and subtracting to the logarithm corresponds to multiplying or dividing the original. Well behaved logarithms imply that two or more values of the original variable should be compared using ratios or percentages.

To actually write it up I have to speak to two audiences. One is me. For me I have to keep it simple — well behaved variables, well-behaved relations between variables when we get to relations. The other audience is a “general public” that will be none to pleased by a statement like “the median number of people per physician is approximately 3.4 in logs base 10”.

So, in order:

1. Transform the data to a well behaved variable.

2. Analyze the transformed data.

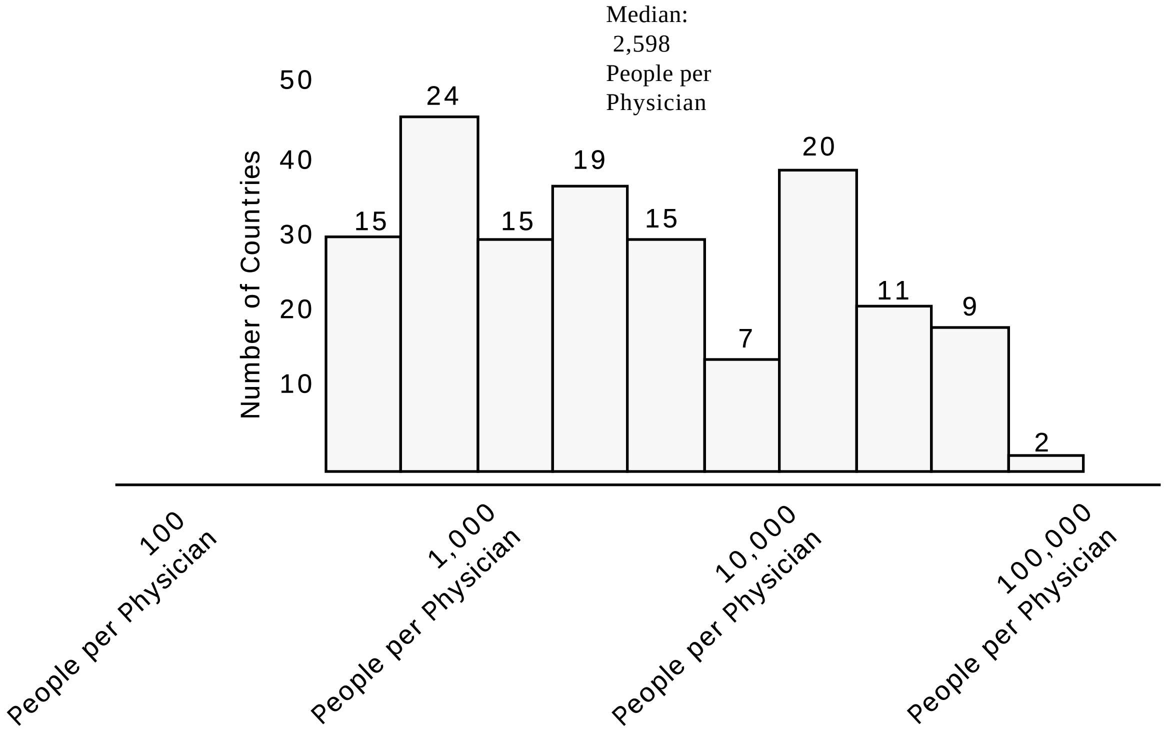
3. Translate the analysis into units of measure that are “friendly” to a non-technical consumer of the data.

For example, using physicians per persons, using logarithms and the rank order statistics here is a brief description of the facts.

Using the logarithm of the number of people per physician, in 1970 the typical country showed a median of 3.41. For example, Mauritania, Saudi Arabia, and Iraq were all close to the median value. The shape of the distribution, Figure __, shows a large range of values but no clear evidence of polarization into two distinct groups, as for example, rich and poor, with nothing in between. Nevertheless the middle fifty per cent of the distribution shows a large range from 2.94 to 4.11. While the full range extends from 2.51 to 4.92, even at the extremes, none are so low or so high as to suggest a sharp differentiation making some of the countries radically different from the rest.	Transform the data Central value Examples Implicit recognition of the shape. Range of typical values Full Range

And now translating

In 1970 the typical country showed a median of approximately 2,600 people per physician. For example, Mauritania, Saudi Arabia, and Iraq were all close to the median value. The shape of the distribution, Figure __, shows a large range of values but no clear evidence of polarization into two distinct groups, as for example, rich and poor, with nothing in between. Nevertheless the middle fifty per cent of the distribution includes more than ten fold contrasts, from fewer than 100 people per physician to more than 1,000 people per physician. The full range extends from an extreme of only 300 people per physician (United Arab Emirates) to an extreme of 80,000 people per physician (Ethiopia), a 300 fold contrast from the most physician intensive to the least physician intensive society. Even at the extremes, none are so low or so high as to suggest a sharp differentiation making some of the countries radically different from the rest.	Invert the transform and translate. The median and quartiles are easy to translate, because the median country is the median country regardless of the unit of measure. The Figure should be translated by re-labelling the x axis in physicians per person even while the shape is computed using log physicians per person. Use words suggesting multiplication (ten fold) because “plus and minus” in terms of logs (original analysis), corresponds to multiplication in terms of people per physician.



And in least square statistics:

Using the logarithm of the number of people per physician, in 1970 the typical country showed a mean of 3.52. For example, the Philippines, Syria, and Honduras were all close to the median value. The shape of the distribution, Figure __, shows a large range of values but no clear evidence of polarization into two distinct groups, as for example, rich and poor, with nothing in between. Nevertheless the variation is large with a standard deviation of .64. The full range extends from 2.51 to 4.92, with physician-poor Equatorial Guinea and Ethiopia standing approximately two standard deviations away from the mean at one end of the distribution at the extremes, none are so low or so high as to suggest a sharp differentiation making some of the countries radically different from the rest.	Central Value Interpretation of the shape Reporting and interpreting the standard deviation Range Marking the extremes using two standard deviations

And now translating

In 1970 the data for 138 countries showed a geometric mean of 3,300 people per physician. For example, the Philippines, Syria, and Honduras were all close to the mean value. The shape of the distribution, Figure __, shows a large range of values but no clear evidence of polarization into two distinct groups, as for example, rich and poor, with nothing in between. Nevertheless the variation is large with a standard deviation around the central value that corresponds to a factor of 4.4.

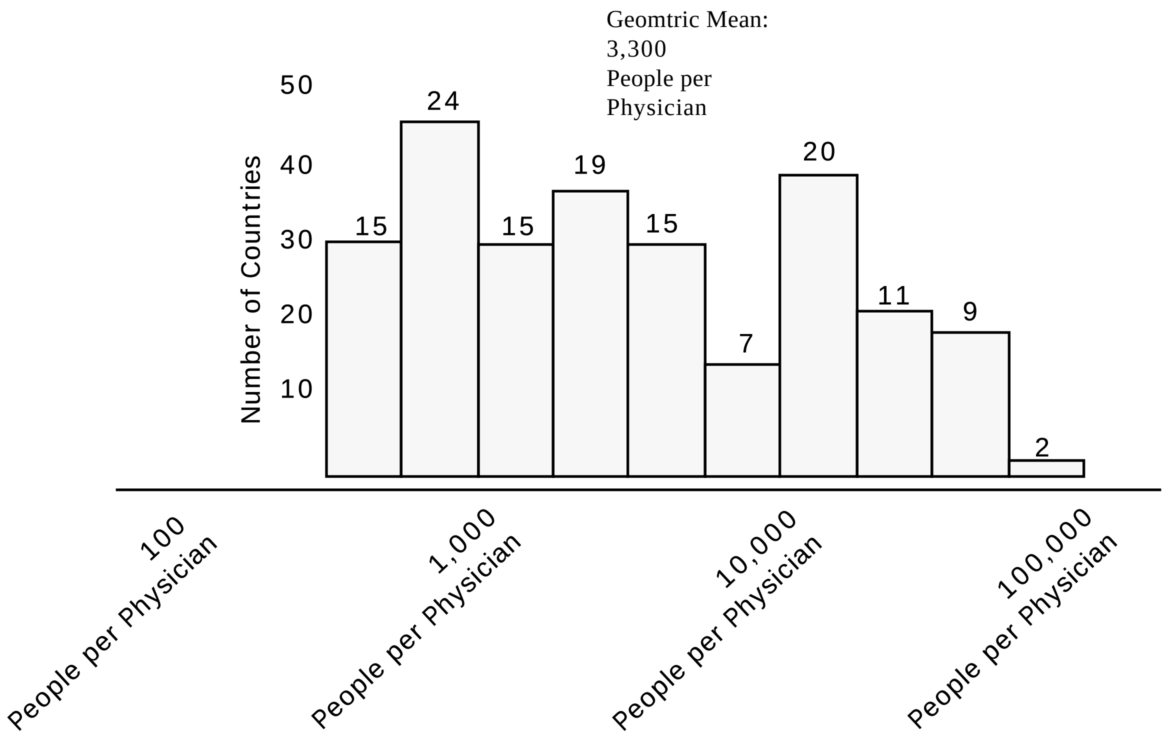
The full range extends from 300 people per physician to 80,000 people per physician, with physician-poor Equatorial Guinea and Ethiopia at extreme values showing more than twenty times the mean value of people per physician.

In order: first transform the data, using logs. Then compute the mean of the logarithms. Then compute the anti-log of the mean. The result is called the geometric mean

The Figure should be translated by re-labelling the x axis in physicians per person even while the shape is computed using log physicians per person.

Use words suggesting multiplication (ten fold) because “plus and minus” in terms of logs (original analysis), corresponds to multiplication in terms of people per physician.

Dodging on my use of plus or minus two standard deviations. The problem is that there is no term in general use for the anti-log of the standard deviation of the log. You would expect it to be called the “geometric standard deviation, but it just does not get named. So, I use it to make an interpretive statement.


Geomtric Mean:
3,300
People per
Physician
50
40
30
20
10
Number of Countries
15
24
15
19
15
7
20
11
9
2
100
People per Physician
1,000
People per Physician
10,000
People per Physician
100,000
People per Physician

Exercise: Consider the data for nations. Using population as the unit of measure, write a brief report summarizing the report, including what is large (and very large). Then, by contrast, use the logarithm of population as the unit of measure and write another brief report. Compare the two? Is China is certainly the largest, by population. But how large? Is it an outlier — so large as to be unrelated to the rest? Or is it merely the largest and not otherwise remarkable?

Exercise: Consider the population data for nations, two different years, and compute the change in population:

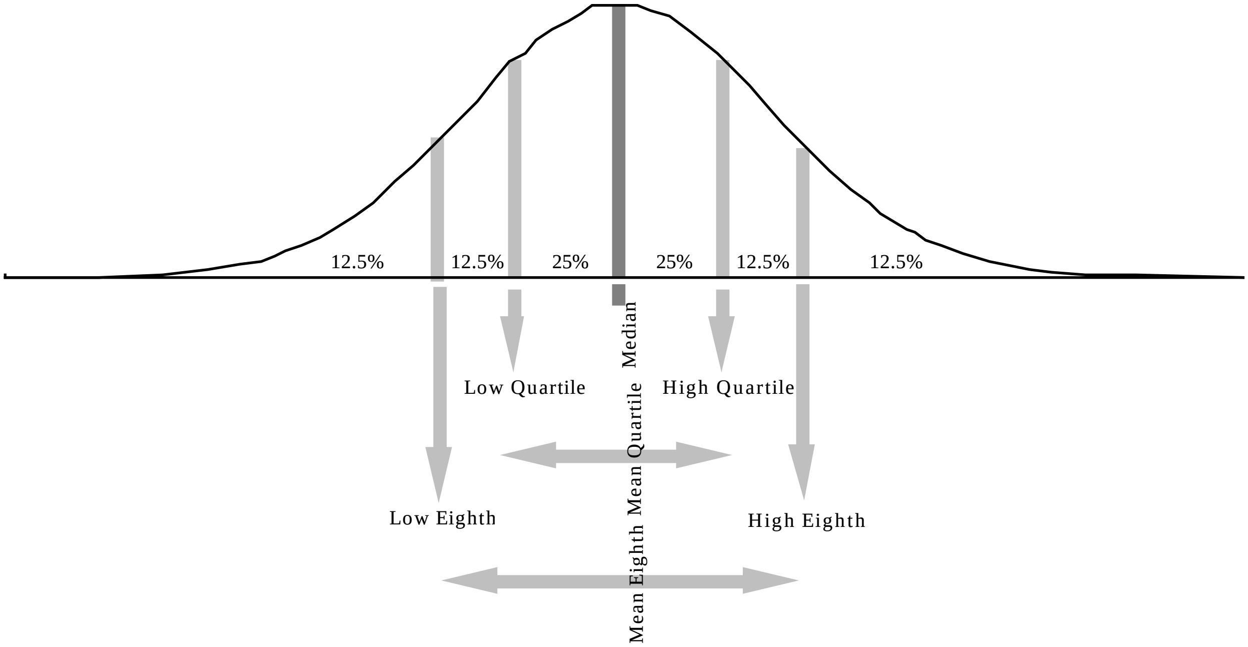
> First, using the nation as the unit of analysis and millions of people as the unit of measure, apply one variable technique, shape of the distribution, measures, and examples, to obtain a brief report of change.
>
> Then, second, using the nation as the unit of analysis and percent of population (first year) as the unit of measure, apply one variable technique, shape of the distribution, measures, and examples, to obtain a brief report of change.

Exercise: As above for GNP (or immigration, or imports v/s imports as a percentage of GNP).

Is it Symmetrical?

Whether or not a variable has a symmetrical distribution is exceedingly important both for descriptive analysis and for more advanced statistical methods. In a simple case there is no need for "high tech" to judge symmetry. Looking back at the people per physician data I feel perfectly competent, on the authority of my eyeball, to look at the picture and assert that the distribution, using people, is not symmetrical. And I feel perfectly competent to look at the second distribution, using logs, is more symmetrical than the first. But for less blatant cases of asymmetry I need a procedure. How should I decide whether data are or are not symmetrical?

The trick is to return to the picture of symmetry and put some numbers on what the eyeball "sees" and identifies as symmetry.



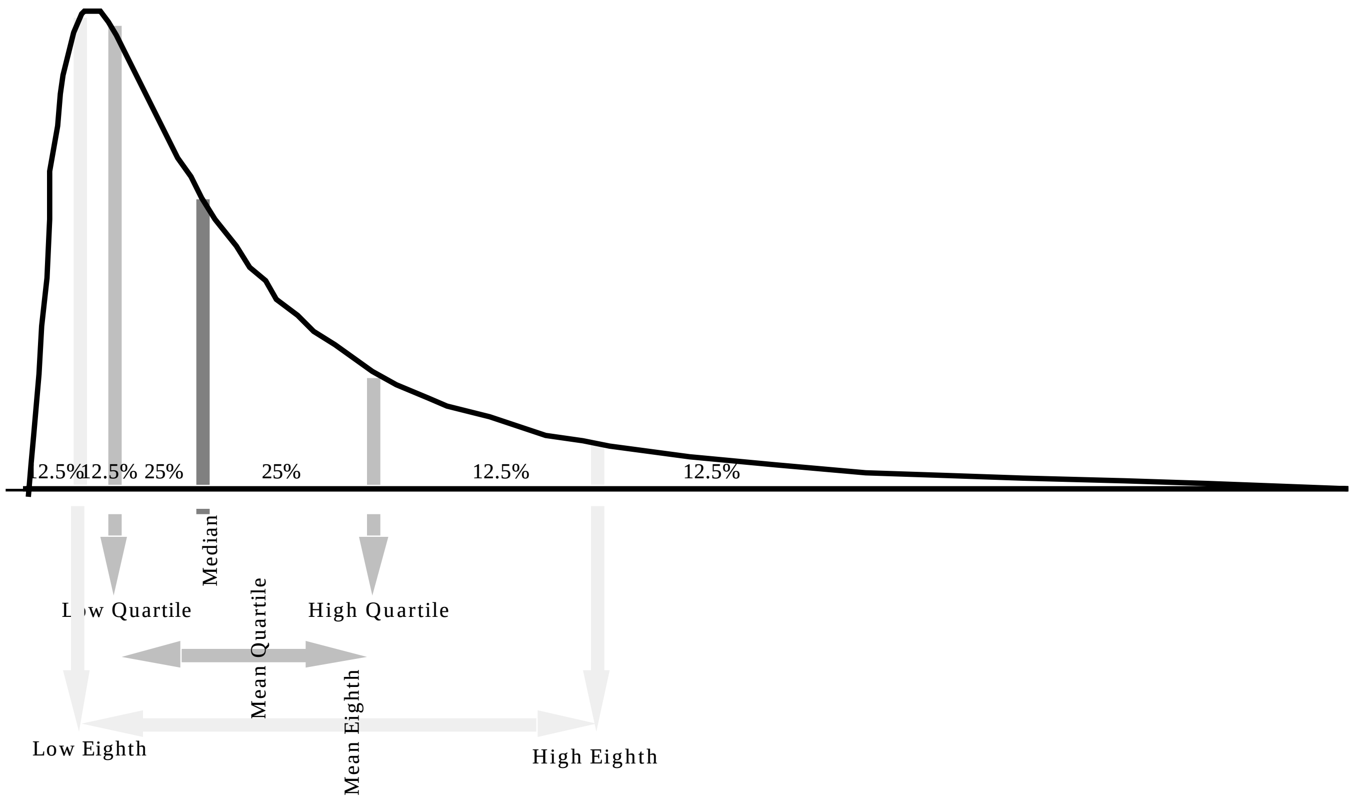
If the distribution is symmetrical, then the quartiles will be located, symmetrically, at equal distances from the median. And, therefore, if the distribution is symmetrical then the point exactly half way between the two quartiles will be equal to the median.

If symmetry, *then* mid-quartile = median

That's easy enough to test: You simply compute the mean quartile and compare it to the median. But generally, two numbers computed from data are rarely equal, they do not match precisely and out to infinite numbers of decimal digits. So we need a test that is a little more clever. For that purpose, following Tukey's Exploratory Data Analysis, compute two more numbers, the two "eighths" and compute the "mid-eighth". Defining terms: As the two quartiles mark the two outer quarters of the distribution, the two eighths mark the two outer eighths of the distribution. And the mid eighth is the point midway between the two eighths. And again, if the distribution is symmetrical then the mid eighth will be equal to the median.

If symmetry, then mid-eighth = median

Now I can get a practical test of symmetry, referring to the asymmetrical distribution in Figure 2: In practice, if there is a trend among the three numbers, from the median to the mid-quartile to the mid eighth, then there is evidence of asymmetry. If the mid-eighth is greater than the mid quartile and the mid quartile is greater than the median, then the distribution is asymmetrical with a tail to the right. If the mid-eighth is less than the mid quartile and the mid quartile is less than the median, then the distribution is asymmetrical with a tail to the left. And if there is no trend, then the distribution is symmetrical. Or — to be very precise (using a double negative): If there is no trend, then there is no evidence of asymmetry.



If you want greater certainty, then you continue the investigation: Adding the mid-sixteenth, the mid-thirty-second … as much as your data will allow.

Defining the "eighths"

To be sure that there is no ambiguity let me specify the step by step computation for the eighths: We find them by mimicking the procedures that have already been used to define the median and the quartiles. Recall that for the fifty-fifty split,

n = number of values in the data

m = location of median =(n+1)/2

And, to repeat, if the result is a whole number then the number of values that are greater than or equal to the median is m If the result is a whole number, then the mth value, in rank order, is the median. If the result is a fraction, then m lies between two values whose mean is the median.

For the quartiles, splitting off twenty-five percent at each end, we compute *m* which is the integer part of m (lopping off the fraction if there is one) and use it to compute the locations of the quartiles

m = number of values greater than or equal to the median

q = location of quartiles = (m+1)/2

Mimicking the logic for the median: if the result, q, is a whole number then the two q-th values, in order from each end of the distribution, are the quartiles. If the result is a fraction then the m-th value at each end lies between two values whose mean is the quartile

are found by counting in q values from each end of the data

identifies the location, then the number of values that are greater than or equal to the median is the integer part of m, *m*.

And now for the eighths, splitting off twelve and one-half percent at each end, we compute q which is the integer part of q (lopping off the fraction if there is one) and use it to compute the locations of the eighths.

q = number of values greater than or equal to the quartile

e = location of the eighths = (*q*+1)/2

If the result, e, is a whole number then the two e-th values, in order from each end of the distribution, are the eighths. If the result is a fraction then the e-th value at each end lies between two values whose mean is the eighths.

Working with the 100 observations of the 10 gram weight, shown in rank order in Table 1, n = 100. So

n = 100

m = (n+1)/2 = (100+1)/2 = 50.5

The median is the mean of the 50-th and 51-st values, median = (9.999596+9.999596)/2 = 9.999596

Then *m* is the integer part of m:

m = 50

q = (*m*+1)/2 = (50+1)/2 =25.5

The high quartile is the mean of the 25th and 26th values in rank order from the high end, Q+ = (9.999599+9.999599)/2 = 9.999599. And the low quartile is the mean of the 25th and 26th values in rank order from the low end, Q- = (9.999593+9.999593)/2 = 9.999593.

Then *q* is the integer part of q:

q = 25

e = (*q*+1)/2 = (25+1)/2 =13

The high eighth is the 13th value in rank order from the high end, E+ = 9.999601. And the low eight is the 13 value in rank from the low end, Q- = 9.999590.

Rank High to Low	Rank Low to High	Item	Weight in Grams
1	100	94	9.999625
2	99	63	9.999608
3	98	85	9.999607
4	97	26	9.999603
5	96	11	9.999602
6	95	97	9.999602
7	94	4	9.999601
8	93	16	9.999601
9	92	22	9.999601
10	91	23	9.999601
11	90	25	9.999601
12	89	29	9.999601
13	88	43	9.999601
14	87	2	9.999600
15	86	17	9.999600
16	85	32	9.999600
17	84	74	9.999600
18	83	7	9.999599
19	82	9	9.999599
20	81	15	9.999599
21	80	18	9.999599
22	79	28	9.999599
23	78	30	9.999599
24	77	34	9.999599
25	76	59	9.999599
26	75	77	9.999599
27	74	83	9.999599
28	73	90	9.999599
29	72	91	9.999599
30	71	5	9.999598
31	70	14	9.999598
32	69	20	9.999598
33	68	24	9.999598
34	67	39	9.999598
35	66	44	9.999598
36	65	50	9.999598
37	64	60	9.999598
38	63	8	9.999597
39	62	10	9.999597
40	61	12	9.999597
41	60	31	9.999597
42	59	67	9.999597
43	58	99	9.999597
44	57	49	9.999596
45	56	51	9.999596
46	55	61	9.999596
47	54	66	9.999596
48	53	69	9.999596
49	52	76	9.999596
50	51	78	9.999596
51	50	89	9.999596
52	49	100	9.999596
53	48	19	9.999595
54	47	40	9.999595
55	46	41	9.999595
56	45	54	9.999595
57	44	62	9.999595
58	43	3	9.999594
59	42	6	9.999594
60	41	37	9.999594
61	40	38	9.999594
62	39	46	9.999594
63	38	52	9.999594
64	37	65	9.999594
65	36	72	9.999594
66	35	80	9.999594
67	34	82	9.999594
68	33	96	9.999594
69	32	98	9.999594
70	31	13	9.999593
71	30	27	9.999593
72	29	35	9.999593
73	28	45	9.999593
74	27	53	9.999593
75	26	64	9.999593
76	25	70	9.999593
77	24	92	9.999593
78	23	21	9.999592
79	22	68	9.999592
80	21	75	9.999592
81	20	79	9.999592
82	19	81	9.999592
83	18	1	9.999591
84	17	42	9.999591
85	16	48	9.999591
86	15	73	9.999591
87	14	95	9.999591
88	13	33	9.999590
89	12	56	9.999590
90	11	57	9.999590
91	10	58	9.999590
92	9	55	9.999589
93	8	71	9.999588
94	7	84	9.999588
95	6	93	9.999588
96	5	47	9.999587
97	4	88	9.999585
98	3	87	9.999582
99	2	36	9.999577
100	1	86	9.999563

Now back to the point, which is to estimate whether or not these data are symmetrical. What we would like is equality: with the median having exactly the same value as the mean quartile and the mean eighth but with real data that is unlikely. What we settle for is a comparison of the median, the mean quartile, and the mean eighth that shows no trend.

For the ten gram weight, what is the evidence:

The median is 9.999596 grams

The mean quartile is (9.999593 + 9.999599) / 2 = 9.999596 grams

The mean eighth is (9.999590 + 9.999601) / 2 = 9.9995955 grams

Reasoning negatively: The numbers do not show clear evidence of asymmetry, so I do not have convincing reason to reject the hypothesis that the measurement errors are described by the hypothesis.

Homework:

1. Pick some easily measured number such as your own pulse (counting for a full 60 seconds to gain precision), or your own blood pressure, or the weight of a coin or the diameter of a coin if you have the equipment. Get at least ten estimates. What is the shape of the distribution for your ten or more estimates?

2. There is a certain ambiguity about the numbers for the ten gram weight: The mean quartile is indistinguishable from the median; the mean eighth is a bit less than the mean quartile. Having more data here, 100 observations, pursue this a fit further: Compute the mean sixteenth and the mean thirty-second. Interpret the whole set of mean value numbers

3. Return to the data for People per Physician, using the logarithm as the unit of measure. Is it symmetrical? Push to the mid sixteenth or further. Is it symmetrical?

Stretching and Shrinking: The Construction of an Interval Scale

One way to understand the concept of a well behaved variable is by the use of another concept employed by data analysts and mathematical modelers. Roughly defined, numerical interval scale must have a correct relation to comparisons among the objects the scale is supposed to represent: If you have measured an object with numbers 1,2,3, then the substance of the differences among the objects must correspond to the differences among the numbers that represent them.

This is a hidden assumption in virtually any numerical procedure applied to data. Consider the mean for example. The mean is so transparent an object that it might seem strange to say that the use of the mean requires certain usually unstated assumptions. That's why I choose it. Recall what a mean is: The mean of a set of numbers is a center that is close to all of the numbers. It is close to them in the sense that it minimizes the squared deviations between the center and the numbers for which it is the center.

There is the key: the deviations. The deviations are a set of intervals: For the first number in the set of data, the deviation is $x_1 - \bar{x}$. That is an interval. For the second number in the set of data, the deviation is $x_2 - \bar{x}$. So when I use the mean, I am assuming that the meanings of these intervals are appropriately represented by the numbers.

When you use the fences to mark out the limits of reasonable variation, you add a number to the high quartile and you subtract a number from the low quartile — which assumes that being so many units above the quartile has a meaning directly comparable to being so many units below the quartile. When you use the standard deviation to mark out limits, again there is an assumption of symmetry, that it is as normal to be one standard deviation above the mean as it is to be one stand deviation below the mean.

Very often these symmetries are not realized as you saw in blatant terms where the boundary for the number of physicians two standard deviations below the mean number of physicians (or below the lower fence) was a negative count — negative physicians — which is ridiculous. That is to say, the moral of the story is that the arithmetic of most data analysis requires interval scales. Without an interval scale even so low tech a computation as the mean is not a valid operation on the numbers. And sometimes the result is not only wrong but obviously wrong as, for example, when it puts the data analyst in the embarrassing position of using numbers that refer to negative people or perhaps negative age or negative income.

In data — as they are presented to the analyst — meaningful numbers are far from guaranteed: For me, counting money as money in hand, the differences between ten dollars in my wallet and twenty dollars and the between ten thousand dollars in my wallet and ten thousand are not the same. From ten to twenty is doubling. From ten thousand to ten thousand and ten the difference is lost in the small change.

But, I have to admit that this statement about unequal intervals is not guaranteed. It depends on context: To an accountant ten dollars is ten dollars. Ten dollars has the same effect on the total (the bottom line) whether it is contributed by an account with little more than ten dollars or one with a great deal more. In this context ten contributes ten to the total wherever it comes from.

If I am measuring traces of a chemical compound, the difference between no trace of the element and one molecule may be extremely important while the difference between one hundred grams of the compound and one hundred and one may have relatively little effect on the conclusions or direction of my research.

For mathematics the differences between numbers may be established by mathematical definition. For the scientist using math to process of assignment of numbers requires some care and depends on context. The use of transformations speaks to the problem of changing the intervals of the scale. The mathematics of these transformations stretches some parts of a scale relative to others, with the consequence that the change of unit can change the behavior of the variable. For example, comparing dollars as the unit of measure to the logarithm of the number of dollars as the unit of measures, note how the logarithm stretches the equal dollar scale at the left in Figure _. Using the dollar as the unit of measure, the four different incomes, $25,000, $50,000, $75,000, and $100,000 are separated by three equal intervals, in dollars.

Re-expressed in logs at the right, the intervals change, stretching the distance between log(25,000) and log(50,000) as compared to the distance between log(50,000) and log(100,000).

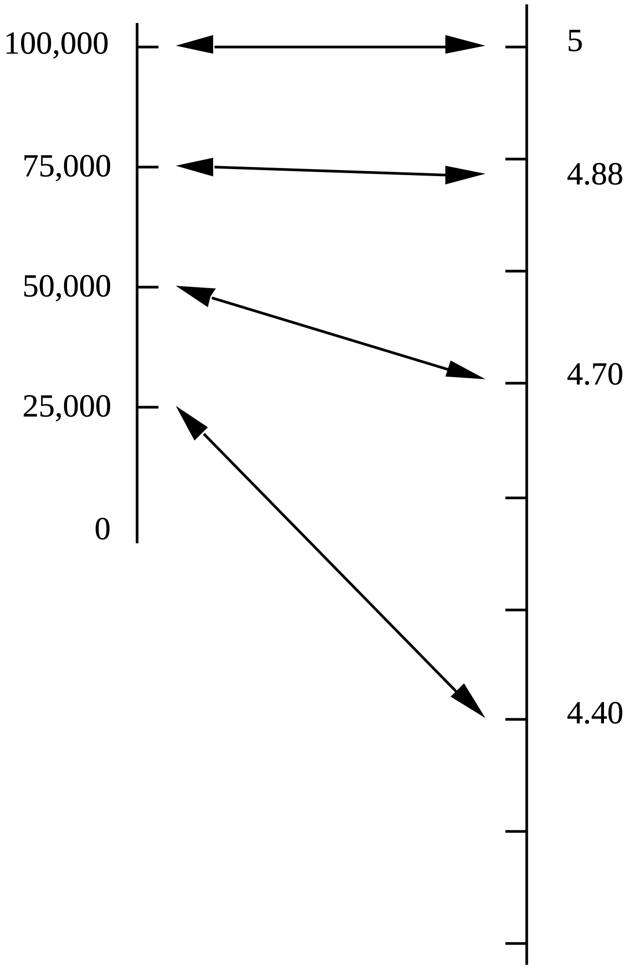


Figure __
Re-Expression of Dollar Values as Logarithmic Values, Using Logarithms Base 10.

Note that the re-expression using logs stretches intervals among small values relative to intervals among the large values.

This “stretching” changes everything: It changes the shape of the distribution, it changes the variation, it changes the relation between one variable and another, and it changes the meaning of the variable. And, in particular, it is capable of transforming a poorly-behaved variable into a well-behaved variable. Here for example is the histogram of the wealth of nations for 19__, first in dollars, and then in log dollars.

Figure: Histograms of gross national products, in dollars and in log dollars.

Exercise

Describe the distribution of gross national products of states of the Western Hemisphere, without logarithms, and with logarithms, in 19__ and 19__ **Get the data**

Exercise: Consider the data for nations. Using population as the unit of measure, write a brief report summarizing the report, including what is large (and very large). Then, by contrast, use the logarithm of population as the unit of measure and write another brief report. Compare the two? Is China is certainly the largest, by population. But how large? Is it an outlier — so large as to be unrelated to the rest? Or is it merely the largest and not otherwise remarkable?

Exercise: Consider the population data for nations, two different years, and compute the change in population:

> First, using the nation as the unit of analysis and millions of people as the unit of measure, apply one variable technique, shape of the distribution, measures, and examples, to obtain a brief report of change.
>
> Then, second, using the nation as the unit of analysis and percent of population (first year) as the unit of measure, apply one variable technique, shape of the distribution, measures, and examples, to obtain a brief report of change.

Exercise: As above for GNP (or immigration, or imports v/s imports as a percentage of GNP).

Transformations

I have tried to convince you by logical argument that things, things out there in the real world, “should” have symmetrical bell-shaped distributions whereas, on the other hand, truth is they do not — not even close. Why? Well, to give you an explanation that tries to salvage both the argument and the reality, consider two hypothetical models of personal income.

Let me imagine a group of 1,000 people, all of whom have an income of $50,000, and watch what happens to them over time. Life can be good and life can be bad: At the end of a year, half of them get a $10,000 increase, half get a $10,000 decrease, half get a $10,000 increase. Now I’ve got 500 people with $40,000 incomes, 500 people with $60,000 incomes.

```
             $50,000
             *
   °                  °
$40,000            $60,000
(500 people)       (500 people)
```

Life goes on and again, half get a $10,000 increase and half get a $10,000 decrease. That gives me 250 people with $30,000, 250 people who dropped to $40,000 and then bounced back to $50,000, 250 more people who rose to $60,000 and then went down to $50,000, and 250 people at $70,000.

$50,000

*

◦ ◦

$40,000 $60,000

◦ ◦ ◦

$30,000 $50,000 $70,000

(250 people) (500 people) (250people)

Let life run on run again, again suppose half go up $10,000 and half go down $10,000

The process seems perfectly ordinary: A few people will got to the top. Some will get to the bottom. The result of their performance, their income distribution, will be the symmetrical result of a symmetrical process.

That’s one look at a hypothetical income process. Here’s another. This time let me start with a group of 1,000 people, all of whom have an income of $50,000, and watch what happens to them over time and then, at the end of a year, half of them get a $10,000 increase, half get a 10% decrease, half get a 10% increase. Now I’ve got 500 people with $40,000 incomes, 500 people with $55,000 incomes.

$50,000

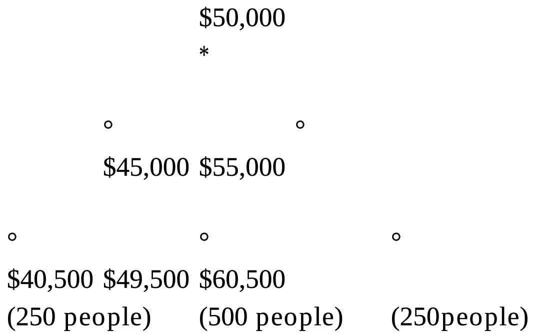
*

◦ ◦

$45,000 $55,000

(500 people) (500 people)

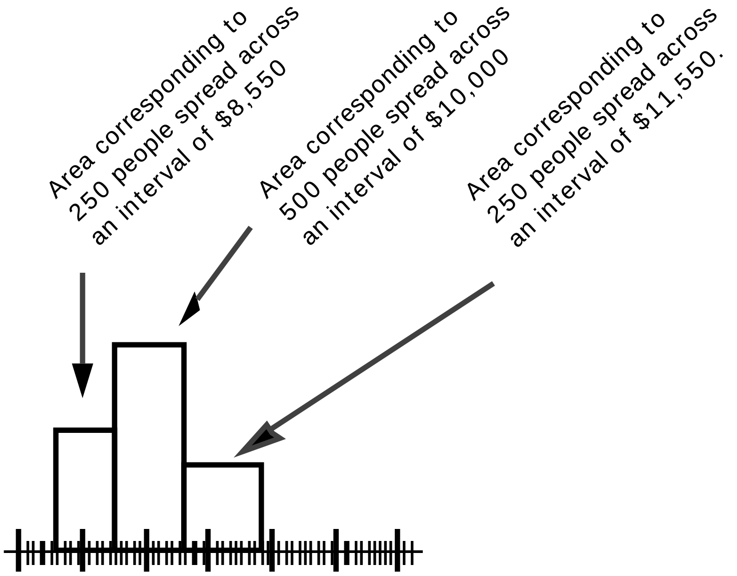
Life goes on and again, half get a 10% increase and half get a 10% decrease. That gives me 250 people with $44,500, 250 people who dropped to $45,000 and then bounced back to $49,500, 250 more people who rose to $55,000 and then went down to $49,500, and 250 people at $60,500.



Again, let life continue for these people, again suppose half go up 10% and half go down 10%. This second process also seems perfectly ordinary: A few people will get to the top,. Some will go to the bottom. If anything this is probably more realistic — these people had income changes that were proportional to the income they already had, some percent up or some percent down. And the second process too has a feel of symmetry about it. But look at the result: These things aren't equally spaced: The gap between the 250 people at the left and the 500 people in the center is $9,000. But the gap between the 500 people at the center and the one at the right is $11,000.

As a result, if we collected these hypothetical data and organized them into a histogram, the histogram would be asymmetrical, skewed to the right.

```
                      $50,000
                         *

              °                   °

              $45,000 $55,000

°             °                   °

     $40,500 $49,500 $60,500

|__250 people_|___500 people_|___250 people_|
$36,450 to   $45,000   to   $55,000   to    $66,550
```



This histogram is only a little bit “off” of symmetry, but it would get worse if I followed it out to allow more and more “bounces” to affect this population, some up and some down. So how do I reconcile this with the privileged place of bell-shaped symmetrical distributions?

The answer is to transform the data. And the reason that that answer is right is because the process itself is not equally spaced in dollars, The process is being performed in percentages. And when you transform the data to a unit of measure that is consonant with the unit in terms of which the process itself is behaving, the result is symmetry.

Data analysts will go one step further, transforming the data using logs rather than percentages. The reason for this is that percentages don’t add up: On an interval scale you want an interval of 1 added to an interval of 1 to add up to an interval of 2, one plus one (should be) equal to 2. But for percentages a 1% increase followed by a second 1% increase does not add up to a 2% increase, not quite. (They combine to a 2.01% increase.) Percentages do not add up. So if you try to draw percentages as an interval scale you get into trouble, more trouble with larger percentages. Percentages are good summary measures because people accept their intuitive meaning. But they get you into trouble if you try to use them in an analysis, even so simple an analysis as a histogram or a stem and leaf.

Logarithms, as compared to percentages “add up”. So we use them where common sense would have us use percentages — because we know that the idea is right but that percentages do not quite do the job.

So for this problem the symmetry of the problem makes itself visible in the picture of the data — using *logarithms*. My people start at log $50,000. Those whose money increases go up from log 50,000 to log 50,000 plus log (1.1): That corresponds to multiplying the $50,000 by 1.1 (increasing it by 10%), except that, using logs, I simply add the logarithm of 1.1.

Those whose money decrease below $50,000 go down from log 50,000 to log 50,000 minus log (1.1): Transformed using logs that is

```
                         log($50,000)
                         *

               o                   o
               log($50,000)-log(1.1)   log($50,000)+log(1.1)

     o                   o                   o
     log($50,000)-2log(1.1)   log($50,000)log($50,000)+2log(1.1)
     250                 500                 250
     people              people              people
```

And now, both the symmetry of the values (in logs) and the symmetry of the counts (in people) are restored.

So, back to the question: How do I reconcile the argument with the facts, the argument that says data should be symmetrical with the fact that data usually are not symmetrical? I reconcile the two by asserting that the data usually *are* symmetrical. But to see the symmetry you have to express the data in units compatible with the process.

If the process is multiplying people incomes or dividing them, then represent the process in logarithms: In logarithms, equal intervals in terms of the logs will correctly represent equal multipliers in terms of the process. And, more interesting: *If* a process looks symmetrical when it is examined in terms of logs, then I infer that the process was symmetrical with respect to multiples.

(Tukey, Chapter 3.) Homework: Look at the distribution of gross national products per capita, by nation. You have the data. And you have the methods for checking for symmetry. So, I ask you, are these data symmetrical in terms of dollars? Are these data symmetrical in terms of log dollars?

And, going further, do the numbers, Tukey style: Using dollars, does the Tukey analysis suggest that some of these nations are not just wealthier than others but different in kind (i.e., beyond the fences)?

Using log dollars, does the Tukey analysis suggests that some of these nations are not just wealthier than others but different in kind (i.e., beyond the fences)? Using different scales — calibrating the Galton board that sorted these nations, but calibrating it in the two different scales, you get two different answers to the last question. Show the two answers. Discuss the discrepancy. And then, practice looking at the world the way I look at it: Argue why someone should take the second interpretation (based on logs) as the correct interpretation. Convince a skeptic.

Thinking About Intervals Using the Tools of Elementary Calculus

One way to understand the transformations is to state a simple question and then use the calculus to derive the answer — which is a transformation.

Here’s the question: I have a variable, x, which changes from case to case. I imagine some cause, c, though I do not assume that I actually know what this cause might be. And I want to look at changes in x related to changes in c.

If I want simple changes in x, there is no problem. I just look at

$$\frac{x(c')-x(c)}{c'-c}$$

And you should recognize from definitions used in elementary calculus, if I look for the limiting form of the relation between x and c as c’ approach c, then this thing becomes the simple derivative for x as a function of c.

$$\frac{dx(c)}{dc}=\lim_{c'\to c}\frac{x(c')-x(c)}{c'-c}$$

Thus the derivative, of the calculus, is a device for expressing simple comparisons.

Now suppose I want to qualify the changes in x by referring them to some other value. For example, suppose I wish to qualify changes in x by comparing them to the size of x itself. Can I find a new variable y such that simple changes in y act like these qualified changes in x?

I can state that question as an equation: Is there a y such that simple changes in y correspond to qualified changes in x?

$$\frac{dy(c)}{dc} = \frac{\frac{dx(c)}{dc}}{x(c)}$$

Fortunately, the equation has a solution. So the answer is “Yes”. The answer uses one of the first differential equations in introductory calculus: Simplifying the equation, it says.

$$dy = \frac{dx}{x}$$

And this differential equation has the solution

$$y = \ln(x)$$

So the answer is, “Yes, use the logarithm of x instead of x itself.”

For the data analyst this has two tactical applications. First, if you want a variable that acts like another variable — but weighted according to the size of the values that are changing, then switch from the original variable to the logarithm of the original variable. (Exercise to the reader: It does not matter which base you use for your logarithms, as long as you are consistent. Prove it.)

Second, the same logic works in reverse: In reverse, suppose I know empirically that the logarithm of a variable is well behaved. I have to ask why: What does it mean when the logarithm of a variable is well-behaved? I answer this question by reverse engineering problem: Knowing that the logarithm is well behaved, what does this tell me about the original variable whose logarithm is well behaved? It tells me that I should be looking at weighted changes, weighted in proportion to size, not simple change.

Generalizing to Other Transformations

Square Root

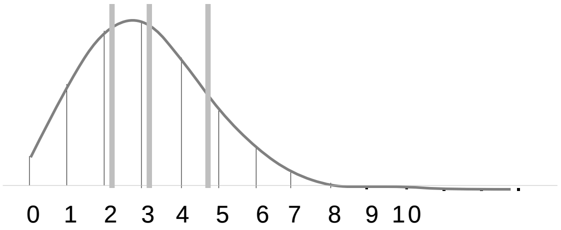
Empirically, counts of objects, tend to have a predictable behavior. Suppose that we are counting the number of people who have incomes between $50,000 and $100,000. Let me suppose that in the general population the number of people in this income category is unknown — some percent of the total. And let me suppose that the data available provides a sample of 1,500 people from the general population. In that sample the number of people with incomes between $50,000 and $100,000 is probably not *exactly* 10%. It is usually a little bit high or a little bit low.

Suppose that another sample of 1,500 becomes available. Again the number of people with incomes between $50,000 and $100,000 will probably not be exactly 10%. It is usually a little bit high or low.

And suppose that yet another sample becomes available. Eventually, with more and more samples, the count will trace a distribution. There will be an average count and there will be a standard deviation for the counts.

So what is the true percentage of the population within this income category? We still don't know. But we can use the mean of the counts computed in these separate samples to estimate the percentage of the general population within this income category?

Both experience and statistical theory tell us certain things about the distribution of counts. Experience tells us that it is likely to have a long tail. And statistical theory tell us that the shape is likely to follow what is called a Poisson distribution. Schematically, it will look something like this.



This is predictable, but it is not “well-behaved” in the specific meaning of that phrase. (It is not symmetrical.)

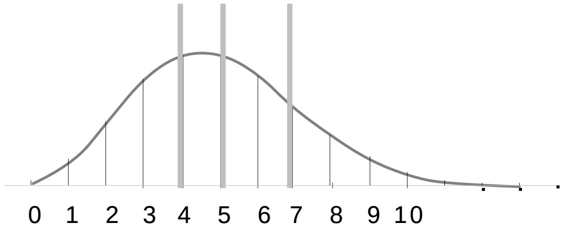
Now suppose I want to compare two counts: Perhaps I have the count of people in this income category in one year and I want to compare it to the count of people in this income category in another year. Or, perhaps I have the count of people in this income category who are also college educated and I want to compare it to the count of people in this income category who have only a high school degree.

How do I compare the counts? The first cut at a comparison is simple: Subtract. That will tell you pretty quickly whether one count is greater than another and how much?

But how big a difference between two counts is a *big* difference? This is not so simple. Suppose that the difference is 2? In the sketch, I’ve assumed that the mean was three for the counts, and sketched-in three vertical lines for the median and the two quartiles. How big is a difference of “2” ? If it is 2 above (if the count was 5), then this is a moderately big difference, slightly more than a quartile away. If it is 2 below (if the count was 1), then this is a big difference, much more than a quartile away.

So is “2” a big difference? It depends, 2 going up is less impressive than 2 going down. “2” at one part of the scale is not the same as “2” at another. That means for us, for those of us who have to interpret these numbers the intervals we are interested are not the intervals in which the data are being measured. That is one of the penalties for trying to work with a variable that is not well-behaved, specifically the penalty for working with a variable that is not symmetrical.

It gets worse. Suppose we have a couple of samples, each of which gives us a number for the second count. Suppose that the mean for these counts for the second group is five. The distribution in this case would look approximately like this



Now, how big is a difference of “2”? The answer is different when this second distribution is used as a reference. So how big is “2”? Well it depends on whether you are going up or going down (asymmetry) and it depends on which distribution you are comparing it to because the variation is different in the two distributions (heteroscedasticity). That is another penalty we pay for failing to work with a well-behaved variable.

So, I want a transformation that is well-behaved. I also know, both empirically and from statistical theory that the standard deviation of a count (or a Poisson distribution) is equal to the square root of its mean. Let me look for a new unit of measure whose simple changes act like changes of counts qualified by comparison to their square roots.

$$dy = \frac{dx}{\sqrt{x}}$$

Solving the equation it tells me to use y as negative two times the square root of x and since the proportionality will not affect the behavior of the result I will use simply y equals the square root of x.

$$y \quad \sqrt{x}$$

So, with counts, try the square root transformation. If you want a variable that acts like another variable — but weighted according to the square root of the values that are changing, then switch from the original variable to the square root of the original variable. (Exercise to the reader: It does not matter whether you use $y=-2\sqrt{x}$ which is the solution to the equation or change the constant of proportionality to use $y=\sqrt{x}$, as long as you are consistent. Prove that if the transformation that is proportional to the square root gives you a unit of measure that is well behaved, then the simple square root itself will also be well behaved of these square root transformations is well-behaved.)

And in reverse, what does it mean when the square root of a variable is well-behaved? I answer this question by reverse engineering problem: Knowing that the square root is well behaved, I should be think that changes of the original variable had to be weighted in proportion to their square roots. So, the original variable is acting like a count.

Postscript on More General Transformations

The logic of this equation can lead to less commonly used transformations. The logic can lead to the inverse, but it is simpler to think of the inverse directly: The inverse of physicians per person is persons per physician. The inverse of time to completion(e.g., the time it takes a runner to complete a mile) is velocity: The inverse of 4 minutes per mile is 15 miles per hour.

The cases we have looked at have had a meaningful minimum at one end: zero people, zero doctors, zero counts, zero velocity. Another type of variable has a meaningful boundary at both ends. For example, what percent of a population is literate? The number is guaranteed to be bounded by 0 at one end and by 100 at the other. So you might wish to count a change from 1 percent literate to 2 percent literate to be a big change, doubling the literacy. By comparison changing the literacy from 50 percent literate to 51 percent literate is probably of little (relatively little) importance. By comparison again, changing the literacy rate from 98 percent to 99 percent is a difficult step, halving the number of illiterates.

By analogy, the equation for logs is comparing x to its lower bound.

$$dy = \frac{dx}{x - \text{lower bound}}$$

Where there are two bounds, the equation becomes

$$dy = \frac{dx}{(x - \text{lower bound})(\text{upper bound } - x)}$$

and the solution becomes

$$y = \log(x - lower\ \ bound) - \log(upper\ \ bound - x)$$

with percentages

$$y = \log(x) - \log(100 - x)$$

and with probabilities

$$y = \log(x) - \log(1 - x)$$

This is useful for data which have either mathematical limits, like percentages and probabilities or systemic limits where “no” production establishes a lower bound and the capacity of a system determines an upper bound.

Why Symmetry?

For a few pages I would like to step aside from the direct business of analyzing data to address the question, "Why symmetry?" Well behaved variables are the key to data analysis that is likely to pay off, as compared to data analysis that churns the numbers a bit without much hope of getting beyond the obvious of means and variation. More sophisticated statistical techniques often condition their results on a premise that the data were well behaved to begin with. Picking the first criterion, what look for symmetry?

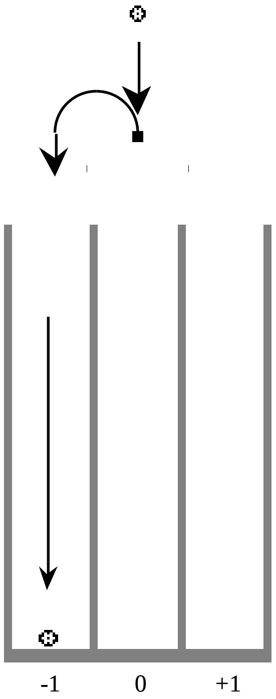
Here I will offer two answers to that question. It is abundantly clear that almost all data will show variation. But why? Two of the possible reasons for variation are error and complexity. I'm going to show you two analogies, one for error, one for complexity. And you will note that each of them leads me to expect symmetry. So beginning with error: Why symmetry?

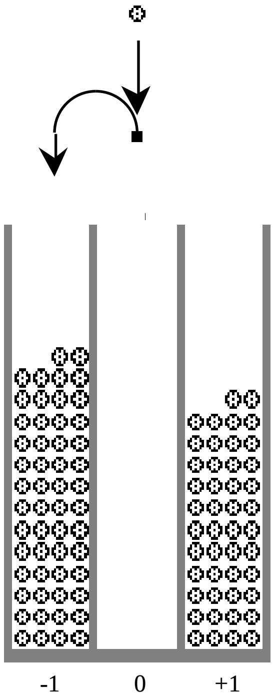
Error
(The Galton Board)

I want you to consider a mechanical model for measurement error: Even the simple process of determining the weight of an object requires a process. And that leads to error. Objects have to be weighed. If the scale is a balance beam then the balance beam has counter weights, and the counter weights themselves have to be measured. A balance beam has a pivot, and the pivot has to be perfectly shaped, which it never is. And the pivot has to be perfectly placed, which it never is. We have to be careful that there is no dirt in the scale, no dust in the pans, and so forth — scales make errors. If a scale has springs in it, then springs have slight irregularities. They are influenced by temperature, and corrosion, not to speak of the fact that we have to know a few laws or a few practical tricks for translating the length of the springs in the scale into numbers that describe the weight of objects.

So measurement introduces errors. The interesting question is not whether or not measurements will include error — the question is, what will the distribution of estimates look like, *including* the error? To answer that question I need a hypothesis. The usual hypothesis is to suppose that error, however it is introduced, is unbiased: The error has a 50/50 chance of increasing the apparent value by some amount and an equal chance of decreasing it by the same amount.

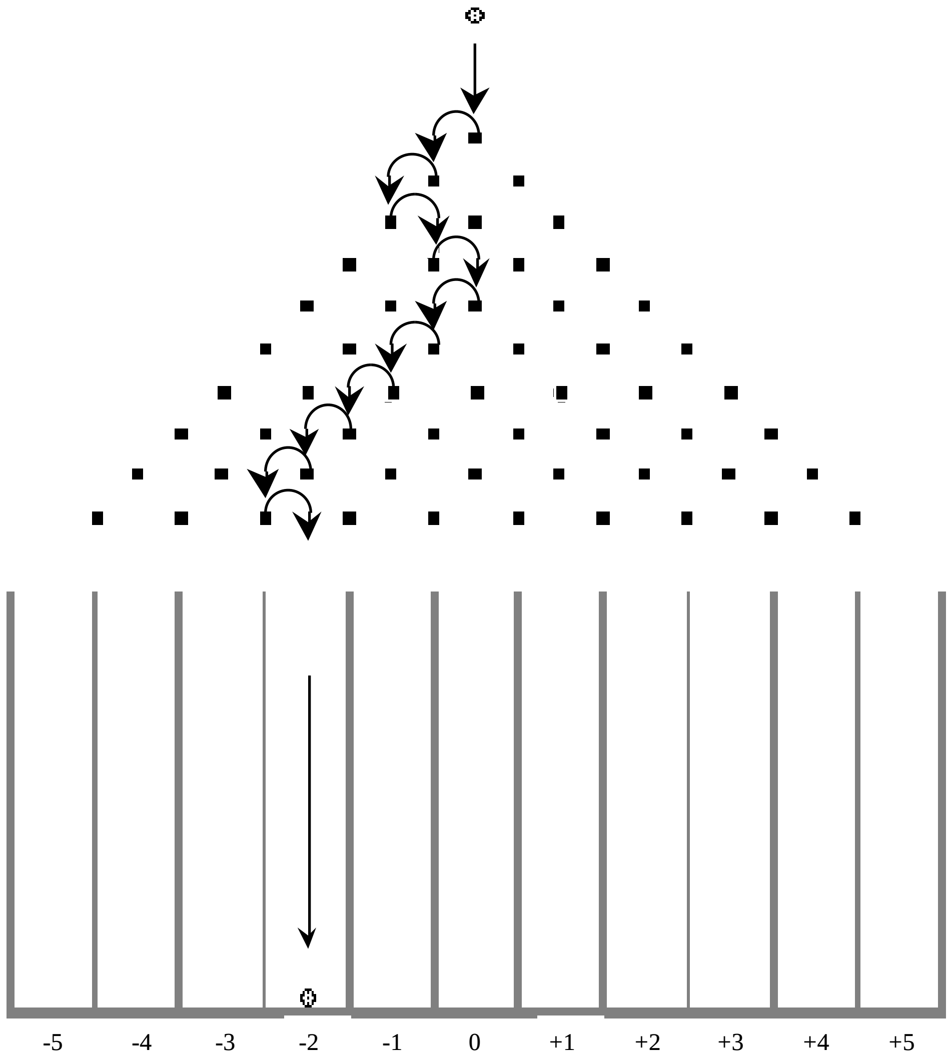
We can duplicate this hypothesis mechanically by imaging that we drop a ball in the direction of a slot that represents the right answer. But, before the ball can fall into the slot it encounters an obstacle that gives the ball a bounce, displacing the ball to the left or to the right of the direction that represents the right answer.



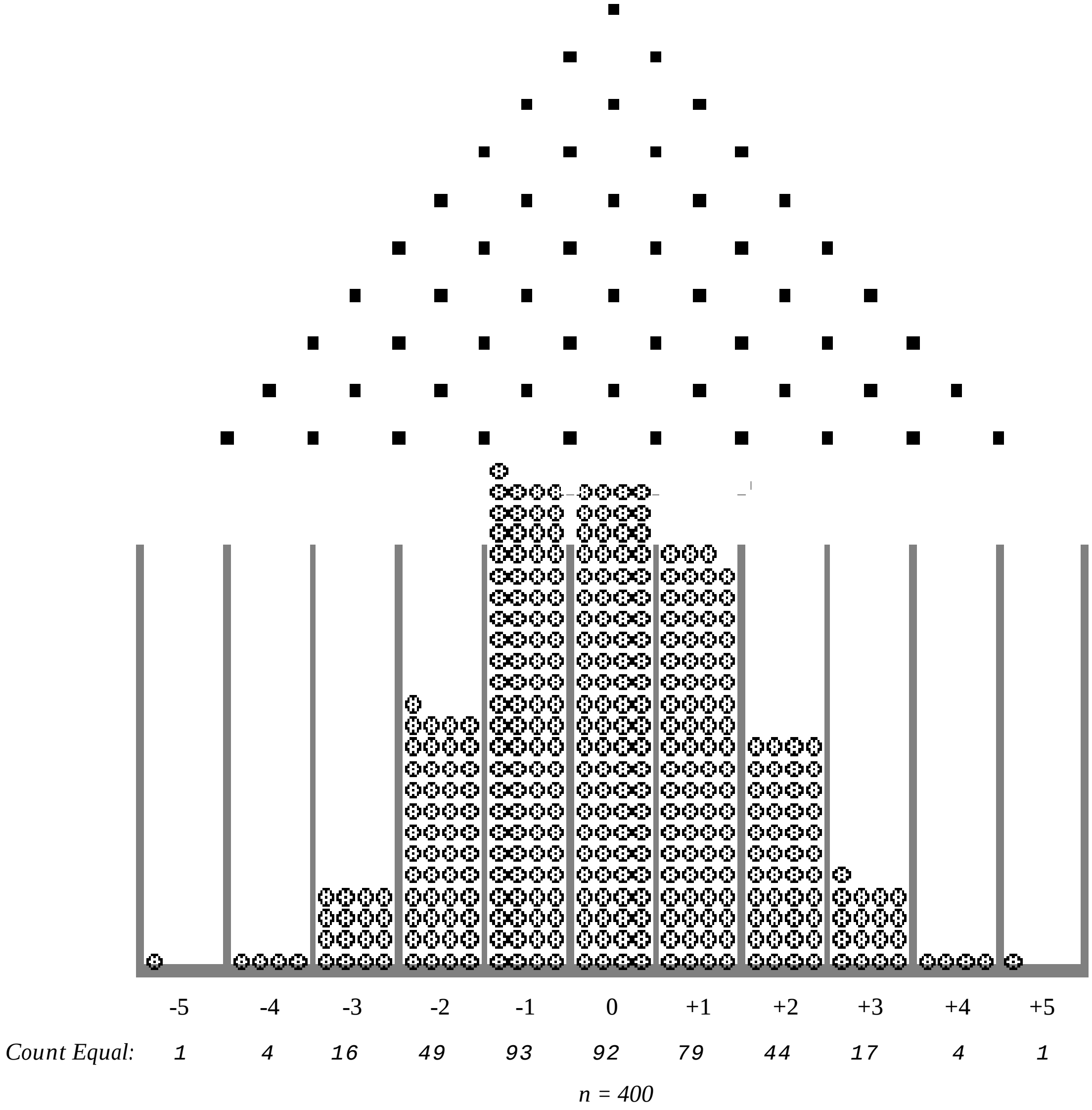


If we make several estimates of the quantity, then the process will repeat itself, making errors, some of the estimates will be less than the correct value, some will be larger than the correct value, and the result will be a distribution of estimates that is “more or less” symmetrical with respect to the correct value — “more or less” because the estimates will tend toward a fifty / fifty split but may not come out exactly fifty / fifty.

The hypothesis further states that a measurement process may include not one but several small errors, each of which has the same effect on the estimate — deflecting the estimate in a direction that makes it high, or deflecting the estimate in a direction which would make it low by the same amount. I can duplicate this process mechanically by imagining a sequence of obstacles standing in the way of my bouncing ball, each one displacing the estimate, each one making the value a little smaller or a little larger.



And if we make several estimates of the quantity, the process will repeat its self, some of the estimates will be small, some will be large, and the result will approximate a symmetrical bell shaped distribution (called the "binomial distribution").



This is the standard hypothesis for errors introduced in the process of measurement. The exact specification of the model might be carried further — how large are the "bounces", what is the width of the slots, how many layers of obstacles are there, and so forth. But for most purposes we are more interested in the moral of this story. The moral of the story is that if the variation of values is due to unbiased measurement error, then the distribution of values should be symmetrical and bell shaped.

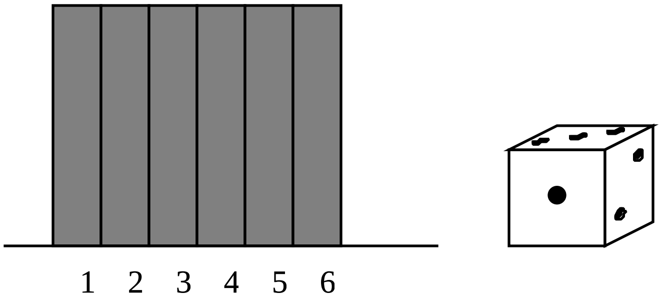
As is frequent in data analysis, the application of this principle is requires that we use it backward: When your data is *not* symmetrical and bell shaped, then you can *not* explain the variation as noise. When the data is not symmetrical and bell shaped, you've got some work to do to explain why not.

rewrite to match 25b

Shape & Number II: Complex Processes

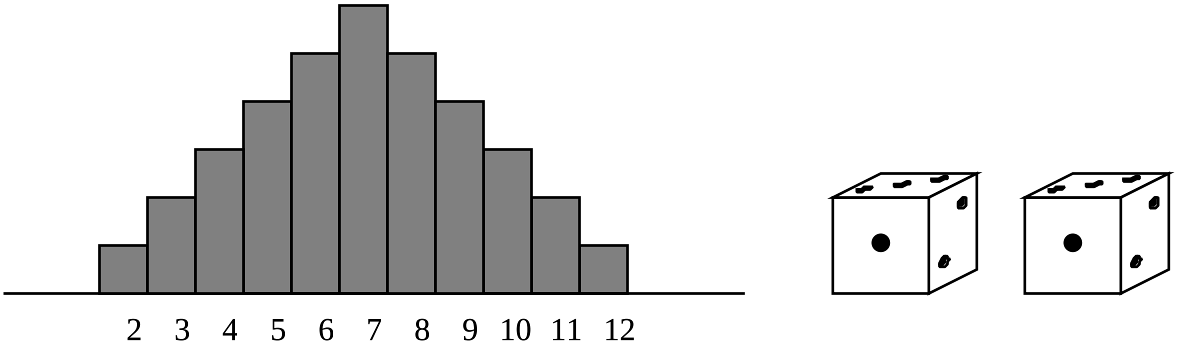
Starting again with the observation that there is something very important about the shape of a distribution, let me introduce another very powerful hypothesis that explains why you *should* expect bell shaped distributions, and why, therefore, it is interesting when that is not what you get. When it is presented mathematically, the principle at work here is known as the Central Limit Theorem, a real mathematical theorem here in the heart of data analysis. What is says, when it is interpreted is that complicated processes will tend to produce bell shaped results.

The central limit theorem is built on the difference between complex events and simple events. Suppose our simple event is analogous to the "process" of throwing a single six-sided die many times. If I threw a single die 6 times or, better, 600 times, I would expect to get something close to an even result, something close to



Theoretical Shape of distribution: One Die, Six Possible Outcomes, 1 through 6, Equal Probabilitie

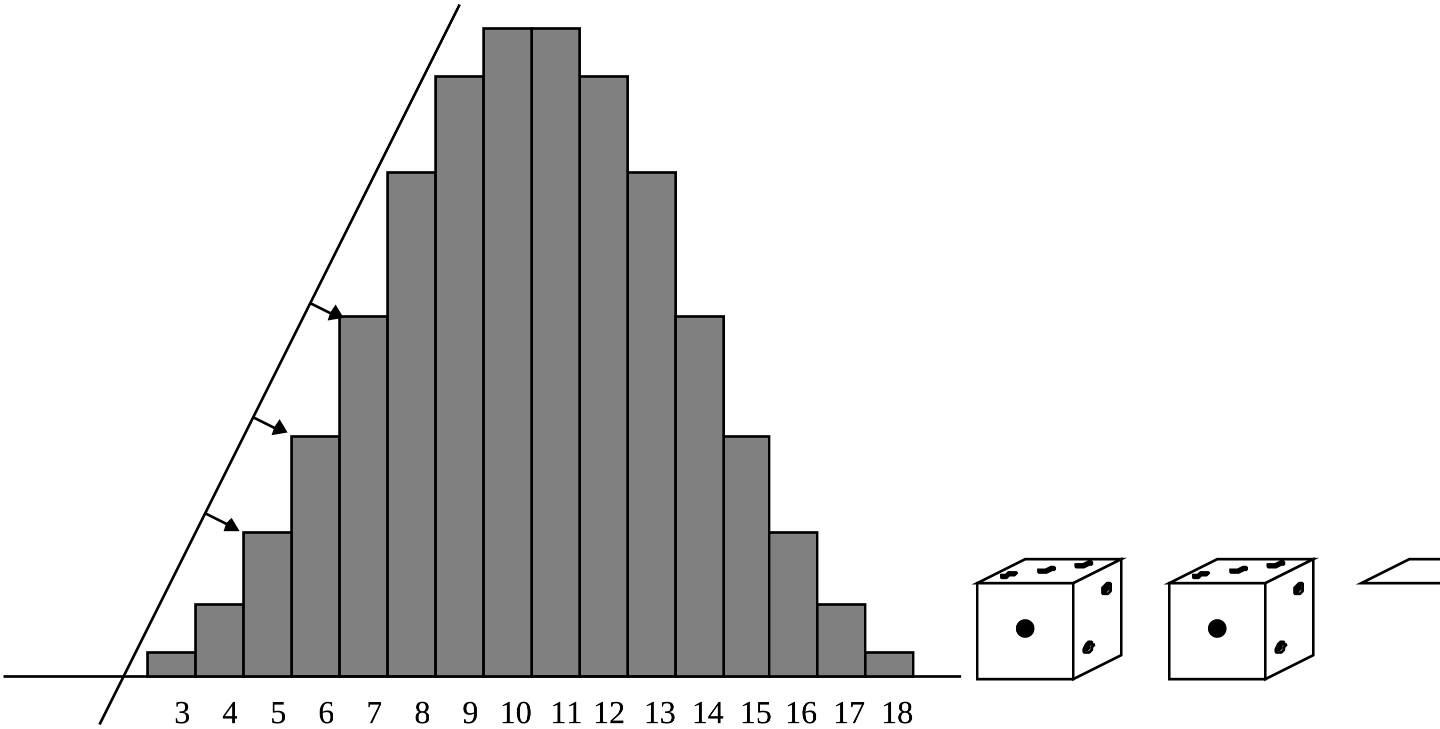
That’s my simple event, and there is the distribution of its results, symmetrical to be sure, but not bell shaped. Now suppose I look at the result of a complicated event, the result of throwing two dice and recording their sum. As is well known, with two dice the results will be uneven: some numbers will be common, others will be rare, specifically, sevens will be relatively common while twos and twelves will be relatively rare. The reason for this is straightforward: There are thirty-six ways that the pair of dice can land and among these six of the thirty-six possibilities will add up to seven (while only one of the thirty six possibilities adds up to two and only one of the thirty six possibilities adds up to twelve). There is 1 way to get a 2, there are 2 ways to get a 3, 3 to get a 4, 4 to get a 5, 5 to get a 6, 6 to get a 7, 5 to get an 8, 4 to get a 9, 3 to get a 10, 2 to get an 11, and 1 to get a 12. So, in 360 throws of the dice or 3,600 throws of the dice I expect a distribution of results something like



Theoretical Shape of distribution: Sum of Two Dice, Eleven Possible Outcomes, 2 through 12, "Triangular" Probabilities

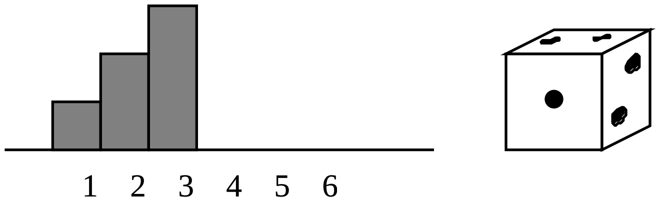
Because most people are familiar with dice there is nothing startling about this behavior of the dice. But it is actually quite remarkable in one way: Specifically, the shape for the composite is not the shape of the things of which it is composed: One die has a flat distribution. But the two dice, together, have a “triangular” distribution.

And with three dice you’ll see a little flaring out, changing shape again toward what is called a bell-shaped or “normal” or “Gaussian” distribution.



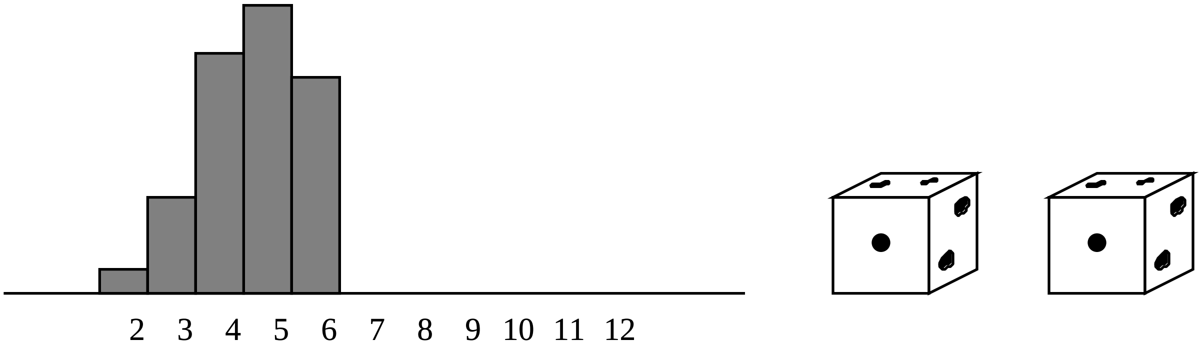
Theoretical Shape of distribution: Sum of Three Dice, Eleven Possible Outcomes, 3 through 18, "Bell-Shaped" Probabilities (Rounded Concave Sides, Rounded Convex Peak)

Dice are so familiar that this little demonstration may fail to convince, so let me try something more extreme. I’m going to change the simple event by taking a marker pen to my set of dice and writing my own numbers on their faces. I’m going to use one 1, two 2’s, and three 3’s. So for 6 or six hundred throws of this simple altered die I expect to get something that is neither symmetrical, nor bell shaped. For one die I should get



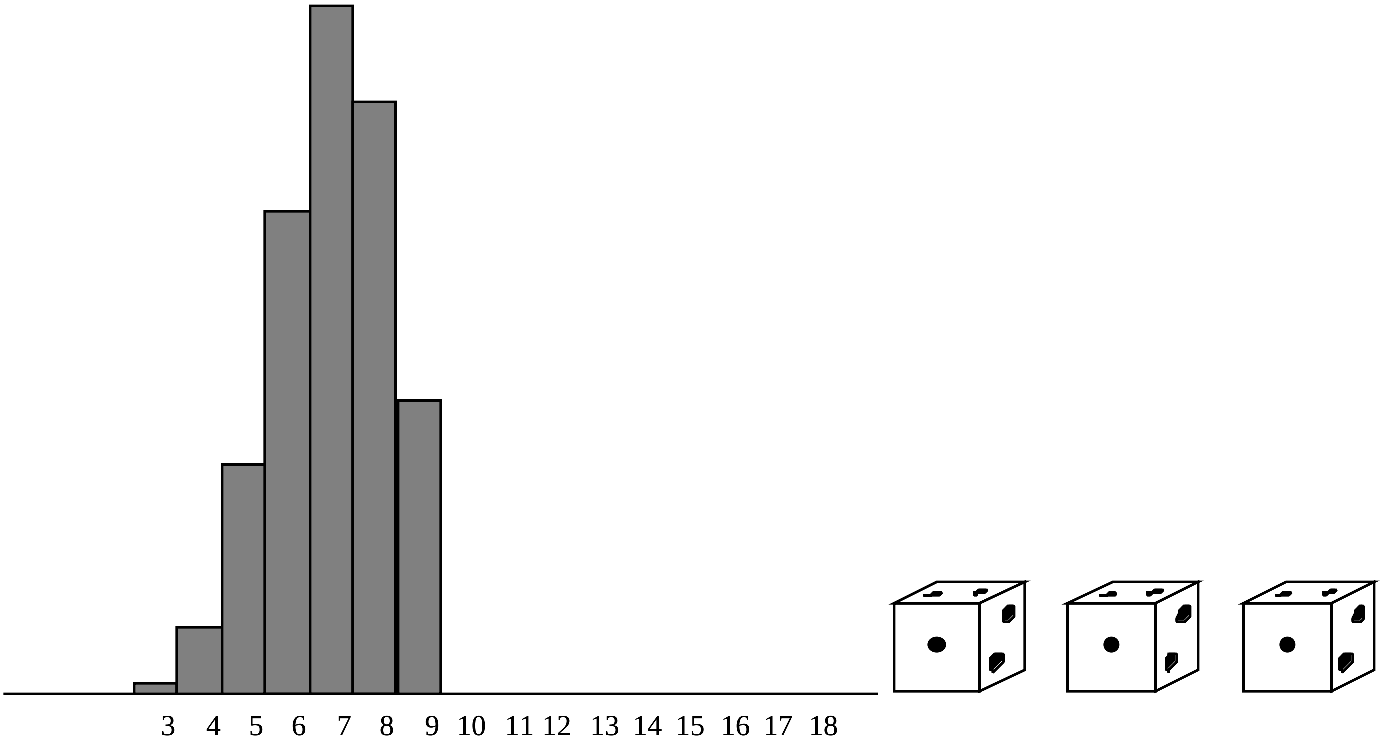
Theoretical Shape of Distribution: One Altered Die with One 1, Two 2's, and Three 3's (Three possible outcomes, 1, 2, or 3 — peak to the right)

There is my simple event, neither bell shaped nor symmetrical. What the Central Limit Theorem says is that the shape of the simple event doesn't matter: The compound event, adding the results of many simple events will always tend to be bell shaped and symmetrical. No matter how weird the simple distribution: combine such events and the result will acquire properties of the Gaussian. Let me put it into action. Throwing two such dice with their individually triangular distributions, what do I get?



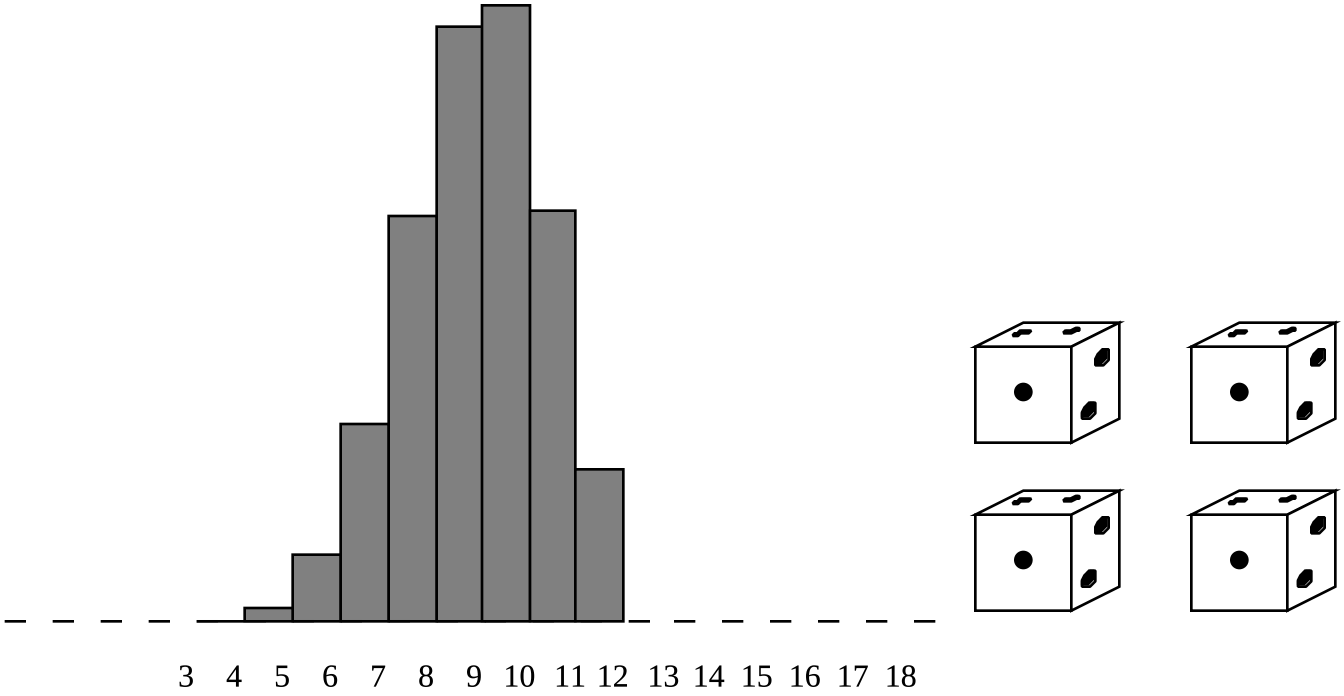
Theoretical Shape of Distribution: Sum of Two Altered Dice, Five Possible Outcomes, 2 through 6 — Peak displaced Toward Center.

Again, the composite shape is not the same as the simple shape. Why? For the same reason that ordinary dice tend to get sevens: With these two dice there are only nine ways of getting the extreme value on the right, while there are twelve ways of getting the peak that is displaced toward the center. Throwing three weird dice, I get.

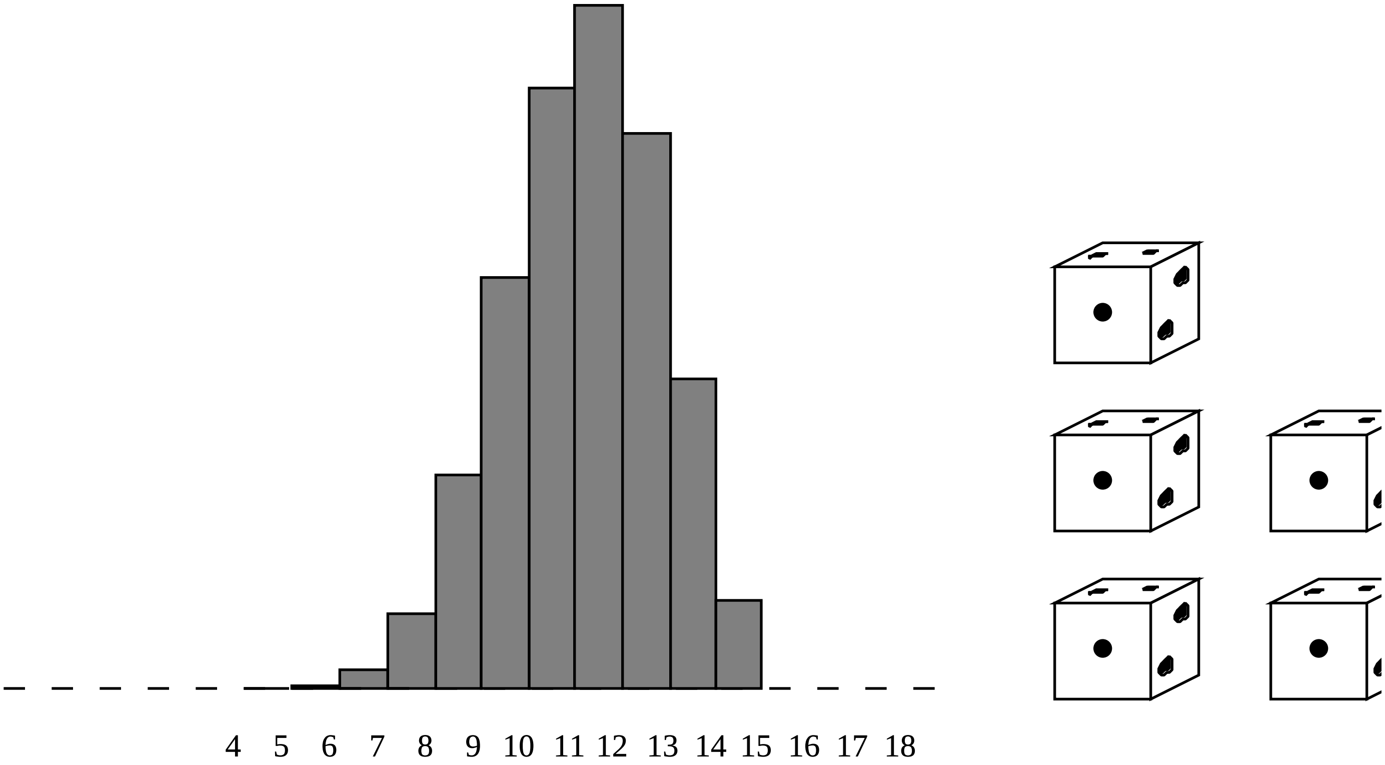


Theoretical Shape of Distribution: Sum of Three Altered Dice, Seven Possible Outcomes, 3 through 9, More Symmetrical (More Bell Shaped: Symmetrical, Rounded Concave Sides, Rounded Convex Peak.

And for four dice and five dice I get



Theoretical Shape of distribution: Sum of Four Altered Dice, Nine Possible Outcomes, 4 through 12, More Symmetrical (More Bell Shaped: Symmetrical, Rounded Concave Sides, Rounded Convex Peak)



Theoretical Shape of distribution: Sum of Five Altered Dice, Eleven Possible Outcomes, 5 through 15, More Symmetrical (More Bell Shaped: Symmetrical, Rounded Concave Sides, Rounded Convex Peak)

The shape is still not bell shaped and symmetrical, but it bears little resemblance to the shape shown by the simple event and it is definitely changing. With more dice, compounding more simple events, the Central Limit Proves that the result will get ever closer to being bell shaped and symmetrical. The Central Limit Theorem as mathematics is more precise than that, defining exactly which properties of these distributions become like the properties of the Gaussian distribution. But the point is that even under extreme circumstances there is good reason to expect symmetrical bell-shaped distributions and to find it interesting when they *don't* happen.

Description Using Logs

Now, using the concept of a well-behaved variable, and using the strategy of re-expression, it's back to the basics of description. But this time we have more tools that can be used for the job, and thus more alternatives that require more thinking. The example will be the size of nations, by population, asking for a brief description of the sizes of nations and how their populations have changed between 1975 and 1990.

Country	***Pop in 1,000's***	
	1975	***1990***
Afghanistan	19,280	15,564
Albania	2,482	3,273
Algeria	16,792	25,337
Andorra		52
Angola	6,394	8,449
Antigua and Barbuda		64
Argentina	25,384	32,291
Armenia		3,357
Aruba		64
Australia	13,809	17,037
Austria	7,538	7,644
Bahamas	200	249
Bahrain	260	520
Bangladesh		113,930
Bangladesh	73,746	
Barbados	245	254
Belgium	9,846	9,909
Belize		220
Benin	3,074	4,674
Bhutan	1,173	1,566
Bolivia	5,410	6,989
Bosnia Herzegovina		4,517
Botswana	691	1,224
Brazil	109,730	152,505
Brunei		372
Bulgaria	8,793	8,934
Burkina		9,078
Burkina		
Burma	31,240	41,277
Burundi	3,765	5,646
Belarus		10,257
Cameroon	6,433	11,092
Canada	22,801	26,538
Cape Verde	292	375
Central African Republ	1,790	2,877
Chad	3,947	5,017
Chile	10,253	13,083
China / People's Republic of		
China / Mainland	838,803	1,133,683
Colombia	25,890	33,076
Comoros	306	460
Congo	1,345	2,242
Costa Rica	1,994	3,033
Croatia		4,686
Cuba	9,481	10,620
Cyprus	673	702
Czechoslovakia	14,793	15,683
Denmark	5,026	5,131
Djibouti		337
Dominica		85
Dominican Republic	5,118	7,241
Ecuador	7,090	10,507
Egypt	37,543	53,212
El Salvador	4,108	5,310
Equatorial Guinea	313	369
Estonia		1,584
Ethiopia	28,134	51,407
Fiji	577	738
Finland	4,652	4,977
France	52,913	56,358
Gabon	521	1,068
Gambia	509	848
Georgia		5,479
Germany		79,123
Germany East	17,127	
Germany West	61,682	63,232
Ghana	9,873	15,130
Greece	8,930	10,028
Grenada	100	84

Guatemala	6,129	9,038
Guinea	4,416	7,269
Guinea-Bissau	525	999
Guyana	791	753
Haiti	4,552	6,142
Honduras	3,037	4,804
Hong Kong	4,225	
Hungary	10,534	10,569
Iceland	216	257
India	613,217	852,667
Indonesia	136,044	190,136
Iran	32,923	57,003
Iraq	11,067	18,782
Ireland	3,131	3,500
Israel	3,417	4,436
Italy	55,023	57,664
Ivory Coast		
Cote d'Ivoire	4,885	12,478
Jamaica	2,029	2,469
Japan	111,120	123,567
Jordan	2,688	3,273
Kampuchea / Cambodi	8,110	6,991
Kenya	13,251	24,342
Kiribati		70
Korea North	15,852	21,412
Korea South	34,663	42,792
Kuwait	1,085	2,124
Kyrgystan		4,394
Laos	3,303	4,024
Latvia		2,695
Lebanon	2,869	3,339
Lesotho	1,148	1,755
Liberia	1,708	2,640
Libya	2,255	4,223
Liechtenstein		28
Lithuania		3,726
Luxembourg	342	384
Madagascar	8,020	11,801
Malawi	4,909	9,197
Malaysia	12,093	17,556
Maldives	120	218
Mali	5,697	8,142
Malta	329	353
Mauritania	1,283	1,935
Mauritius	899	1,072
Mexico	59,204	88,010
Moldova		4,393
Mongolia	1,446	2,187
Morocco	17,504	25,630
Mozambique	9,223	14,539
Namibia		1,453
Nepal	12,572	19,146
Netherlands	13,599	14,936
New Zealand	3,031	3,296
Nicaragua	2,318	3,602
Niger	4,600	7,879
Nigeria	63,049	118,819
Norway	4,007	4,253
Oman	770	1,481
Pakistan	70,560	114,649
Panama	1,678	2,425
Papua New Guinea	2,716	3,823
Paraguay	2,647	4,660
Peru	15,326	21,906
Philippines	44,437	64,404
Poland	33,841	37,777
Portugal	8,762	10,354
Puerto Rico	2,902	
Qatar	90	491
Romania	21,178	23,273
Russia		148,254
Rwanda	4,233	7,609
Saint Kits and Nevis		40
Santa Lucia		150
Saint Vincent and the Grenadines	80	113
San Marino		23
Sao Tome and Principe		125
Saudi Arabia	8,966	17,116
Senegal	4,418	7,714
Serbia		9,883
Seychelles	60	68
Sierra Leone	2,983	4,166
Singapore	2,248	2,721
Somalia	3,170	6,654
South Africa	24,663	39,539
Soviet Union frmr		
Spain	35,433	39,269
Sri Lanka	13,986	17,198
Sudan	18,268	26,245
Suriname	422	397
Swaziland	469	837
Sweden	8,291	8,526
Switzerland	6,535	6,742
Syria	7,259	12,483
Tajikistan		5,342
Taiwan / Republic of China	16,453	20,435
Tanzania	15,388	25,971
Thailand	42,093	56,002
Togo	2,248	3,674
Trinidad and Tobago	1,009	1,271
Tunisia	5,747	8,104
Turkey	39,882	57,285
Turkmenistan		3,658

Tuvalu		9	Venezuela	12,213	19,698
UAR United Arab			Vietnam	43,451	66,171
Emirates	220	2,254	Vietnam North	23,800	
Uganda	11,353	18,016	Vietnam South	19,650	
Ukraine		51,711	Western Somoa	160	186
United Kingdom	56,427	57,366	Yemen		9,746
Upper Volta	6,032		Yemen (Aden)	1,660	
Uruguay	3,108	3,102	Yemen (Sana)	6,668	
USA	213,925	250,410			
USSR	255,038		Yugoslavia	21,322	
			Zaire	24,450	36,613
Uzbekistan		20,569	Zambia	5,004	8,154
Vanuatu		165	Zimbabwe	6,272	10,394

Table 1

Countries of the World: 1975 and 1990 Population

Source, 1975: *World Handbook of Political and Social Indicators, Volume I*, Taylor and Jodice. Original source Labour Force Estimates and Projections, 1950-2000, ILO, Geneva, 1977, and Demographic Yearbook, 1977.. Source, 1990: *Statistical Abstract of the United States*, 1991 Table 1,359, compiled by the U.S. Census Bureau from various original sources.

World Population: The Work

So now, from the beginning: Who, What, Where, ...: The data are from the United Nations and in turn from national sources. How good are the data? Well, the text in the secondary sources I am using, from *World Handbook of Political and Social Indicators* and the *Statistical Abstract of the United States* warns me that standards differ from country to country. For example, some do and some do not count aborigines, nomadic peoples, displaced persons, or refugees. The separate counts are based on varying methods including attempts at complete counts, including samples, including registration censuses based on voting or tax registers. So, the data are a mixed lot. But, the data are also the best I can get — the United Nations sources have attempted to adjust for these inconsistencies. And were I reject these data on population, notwithstanding their blemishes, I would be acting as if there were no data on populations — because I would have rejected the best. So, I'll accept the data, with caution.

Now for a first look at the data, stem and leaf. The national populations range from a low of "9", which is nine thousand, to a high of 1,133,683, which is one billion. If I attempt to break this range into approximately ten equal stems, dividing the range into intervals of 100,000 each (one hundred million each), I will get a mess — I can see that coming by just looking at the counts, without completing the stem and leaf:

<table>
<tr><th colspan="2">Stems</th><th>Leaves</th><th></th></tr>
<tr><td>0-</td><td>100,000</td><td>| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | .....</td><td>170countries</td></tr>
<tr><td>100,000 -</td><td>199,999</td><td>| | | | | | |</td><td>7 countries</td></tr>
<tr><td>200,000 -</td><td>299,999</td><td>|</td><td>1 country</td></tr>
<tr><td>-------</td><td></td><td></td><td></td></tr>
<tr><td>800,000 -</td><td>899999</td><td>|</td><td>1 country</td></tr>
<tr><td>-------</td><td></td><td></td><td></td></tr>
<tr><td>1,000,000 -</td><td>|</td><td>1 country.</td><td></td></tr>
</table>

No point in completing this stem and leaf, I already know what it looks like: Most of the countries are "piled up" at the low end of the scale. There are a few very large countries forming a "tail" at the high end of the distribution.

I could persist with the stem and leaf, changing scales, omitting very large nations, and doing it again. But I'm in a hurry, I'm always in a hurry, so I'll compromise by simply putting the countries in order by size. The printed page, ranked by size, gives me much of what I need from the stem and leaf and, since this is an extremely "skewed" distribution, (very asymmetrical, with a few very large values) it avoids the work of the stem and leaf — which (I already know) is unlikely to pay off with a good-looking stem and leaf.

	Country	***Pop in 1,000's***	
		1975	***1990***
1	Tuvalu		9
2	San Marino		23
3	Liechtenstein		28
4	Saint Kits and Nev		40
5	Andorra		52
6	Antigua and Barbuda		
7	Aruba		64
8	Seychelles	60	68
9	Kiribati		70
10	Grenada	100	84
11	Dominica		85
12	Saint Vincent and	80	113
13	Sao Tome and Princi		125
14	Santa Lucia		150
15	Vanuatu		165
16	Western Somoa	160	186
17	Maldives	120	218
18	Belize		220
19	Bahamas	200	249
20	Barbados	245	254
21	Iceland	216	257
22	Djibouti		337
23	Malta	329	353
24	Equatorial Guine	313	369
25	Brunei		372
26	Cape Verde	292	375
27	Luxembourg	342	384
28	Suriname	422	397
29	Comoros	306	460
30	Qatar	90	491
31	Bahrain	260	520
32	Cyprus	673	702
33	Fiji	577	738
34	Guyana	791	753
35	Swaziland	469	837
36	Gambia	509	848
37	Guinea-Bissau	525	999
38	Gabon	521	1,068
39	Mauritius	899	1,072
40	Botswana	691	1,224
41	Trinidad and T	1,009	1,271
42	Namibia		1,453
43	Oman	770	1,481
44	Bhutan	1,173	1,566
45	Estonia		1,584
46	Lesotho	1,148	1,755
47	Mauritania	1,283	1,935
48	Kuwait	1,085	2,124
49	Mongolia	1,446	2,187
50	Congo	1,345	2,242
51	UAR United Arab	220	2,254
52	Panama	1,678	2,425
53	Jamaica	2,029	2,469
54	Liberia	1,708	2,640
55	Latvia		2,695
56	Singapore	2,248	2,721
57	Central African	1,790	2,877
58	Costa Rica	1,994	3,033
59	Uruguay	3,108	3,102
60	Albania	2,482	3,273
61	Jordan	2,688	3,273
62	New Zealand	3,031	3,296
63	64Lebanon	2,869	3,339
64	Armenia		3,357
65	Ireland	3,131	3,500
66	Nicaragua	2,318	3,602
67	Turkmenistan		3,658
68	Togo	2,248	3,674
69	Lithuania		3,726
70	Papua New Guinea	2,716	3,823
71	Laos	3,303	4,024
72	Sierra Leone	2,983	4,166
73	Libya	2,255	4,223
74	Norway	4,007	4,253
75	Moldova		4,393
76	Kyrgystan		4,394
77	Israel	3,417	4,436
78	Bosnia Herzegovina		4,517
79	Paraguay	2,647	4,660
80	Benin	3,074	4,674
81	Croatia		4,686
82	Honduras	3,037	4,804
83	Finland	4,652	4,977
84	Chad	3,947	5,017
85	Denmark	5,026	5,131
86	El Salvador	4,108	5,310
87	Tajikistan		5,342
88	Georgia		5,479
89	Burundi	3,765	5,646
90	Haiti	4,552	6,142
91	Somalia	3,170	6,654
92	Switzerland	6,535	6,742
93	Bolivia	5,410	6,989
94	Kampuchea / C	8,110	6,991
95	Dominican Rep	5,118	7,241
96	Guinea	4,416	7,269
97	Rwanda	4,233	7,609
98	Austria	7,538	7,644
99	Senegal	4,418	7,714
100	Niger	4,600	7,879
101	Tunisia	5,747	8,104
102	Mali	5,697	8,142
103	Zambia	5,004	8,154
104	Angola	6,394	8,449
105	Sweden	8,291	8,526
106	Bulgaria	8,793	8,934
107	Guatemala	6,129	9,038
108	Burkina		9,078
109	Malawi	4,909	9,197
110	Yemen		9,746

111	Serbia		9,883
112	Belgium	9,846	9,909
113	Greece	8,930	10,028
114	Belarus		10,257
115	Portugal	8,762	10,354
116	Zimbabwe	6,272	10,394
117	Ecuador	7,090	10,507
118	Hungary	10,534	10,569
119	Cuba	9,481	10,620
120	Cameroon	6,433	11,092
121	Madagascar	8,020	11,801
122	Ivory Coast / C	4,885	12,478
123	Syria	7,259	12,483
124	Chile	10,253	13,083
125	Mozambique	9,223	14,539
126	Netherlands	13,599	14,936
127	Ghana	9,873	15,130
128	Afghanistan	19,280	15,564
129	Czechoslovakia	14,793	15,683
130	Australia	13,809	17,037
131	Saudi Arabia	8,966	17,116
132	Sri Lanka	13,986	17,198
133	Malaysia	12,093	17,556
134	Uganda	11,353	18,016
135	Iraq	11,067	18,782
136	Nepal	12,572	19,146
137	Venezuela	12,213	19,698
138	Taiwan / Rep	16,453	20,435
139	Uzbekistan		20,569
140	Korea North	15,852	21,412
141	Peru	15,326	21,906
142	Romania	21,178	23,273
143	Kenya	13,251	24,342
144	Algeria	16,792	25,337
145	Morocco	17,504	25,630
146	Tanzania	15,388	25,971
147	Sudan	18,268	26,245
148	Canada	22,801	26,538
149	Argentina	25,384	32,291
150	Colombia	25,890	33,076
151	Zaire	24,450	36,613
152	Poland	33,841	37,777
153	SPAN	35,433	39,269
154	South Africa	24,663	39,539
155	Burma	31,240	41,277
156	Korea South	34,663	42,792
157	Ethiopia	28,134	51,407
158	Ukraine		51,711
159	Egypt	37,543	53,212
160	Thailand	42,093	56,002
161	France	52,913	56,358
162	Iran	32,923	57,003
163	Turkey	39,882	57,285
164	United Kingdom	56,427	57,366
165	Italy	55,023	57,664
166	Germany West	61,682	63,232
167	Philippines	44,437	64,404
168	Vietnam	43,451	66,171
169	Germany		79,123
170	Mexico	59,204	88,010
171	Bangladesh		113,930
172	Pakistan	70,560	114,649
173	Nigeria	63,049	118,819
174	Japan	111,120	123,567
175	Russia		148,254
176	Brazil	109,730	152,505
177	Indonesia	136,044	190,136
178	USA	213,925	250,410
179	India	613,217	852,667
180	China / Peop	838,803	1,133,683
181	Bangladesh	73,746	
182	Burkina		
183	Germany East	17,127	
184	Hong Kong	4,225	
185	Puerto Rico	2,902	
186	Soviet Union frmr		
187	Upper Volta	6,032	
188	USSR	255,038	
189	Vietnam North	23,800	
190	Vietnam South	19,650	
191	Yemen (Aden)	1,660	
192	Yemen (Sana)	6,668	
193	Yugoslavia	21,322	

Ah, I notice immediately from the rank order on 1990 population, that I don't have a 1990 population for all of theses countries — the list of nations varies from year to year. Checking the names, I see that the change in nations is a result of fusion and fission. Do I attempt to compensate for this, changing the units of the 1975 list to correspond to the units of the 1990 list, standardizing my list? No, at some point that may be called for, but to decide on the "correct" standardization I would have to have a clear purpose in mind (with respect to which I could decide what was “right”.) Having no particularly subtle purpose in mind, beyond description, changing the list now would simply change the data — to no apparent end. So, I’ll continue, acknowledging that it may be difficult to compare the 1975 data to the 1990.

Now, I’m going to stop for a moment: Why? Because I was about to yield to a mindless reflex: I was about to compute averages and measures of variation on everything in sight. That kind of mindless reflex is to be treated with great caution. So, thinking about these data, before I commit to a lot of computation, what do I already know and what do I suspect: Invoking the list of 4 properties for well-behaved variables, I know that population fails criterion #1 and therefore I suspect that population will fail on criterion #4. I know that the distribution is not symmetrical, criterion #1. And therefore, I suspect that the unit of measure is the wrong unit, criterion #4.

1990 population surely fails the first property of a well behaved variable — it is anything but symmetrical. And, ordinarily that would be enough to stop me, but for pedagogical purposes, let me show you the kind of trouble I would get in to if I yielded to mindless reflex.

[Check how Excel defines median and how it uses missing data: I got a different value for the median, using the median built in function, than I got by picking the median out of the rank order.]

	1975	1990	
Sum	4,021,291	5,347,251	Added because I didn’t trust the dif between the means and wanted, theref look at the difference between the sum
Mean	25,944	29,707	**Verify directly**
Standard Deviation (Excel function stdevp)	87,477	111,616	**Verify directly**
Median	5,722	6,398	**Verify**
Low Quartile	1,994	1,699.5	**Verify**
High Quartile	17,127	17,786	**Verify**
Quartile Spread	15,133	16,119.5	

[Looking that over, first look at the means: increasing from 26 million to 30 million, about 16 percent. That seems a little odd, 16 percent increase during fifteen years is approximately one percent a year. That seems low — my memory tells me that the growth rate for world population is about 2 percent per year. Shouldn't the average also be growing by about 2 percent?

Well, still thinking, maybe and maybe not: This is not *world population*, it is average size of nations, which is different. So, maybe. Let me re-assure myself by checking world population, adding up the populations: Ah, 4 billion in 1975 up to 5.3 billion, up about 33%. So, yes, the total seems in line with what I expect, approximately 2 percent per year. That warns me to be careful about the unit: these are nations (actually states). The means may be showing the trace of the breakup of the Soviet Union into smaller countries. O.K. I'm ready to continue.]

First, look at those standard deviations: In both cases they exceed the means, substantially — the standard deviations are more than three times greater than their respective mean. That's strange: If those are the numbers, then those are the numbers, barring numerical error. But still, think of what those numbers are supposed to mean: They are supposed to represent — and put a number on — what you saw in the picture. The "standard deviation" is supposed to describe standard or typical average variation around the mean but these numbers are much too large for that purpose: For 1975, 150 of the 155 countries are *less than* one standard deviation away from the mean.

Putting it another way, remember that we will use a standard deviation to describe the middle range of data. But it just doesn't do the job with this picture: By the numbers, the middle range of data would be between – 61,533 (minus 62 million) and + 113,421 (plus 113 million) — calculating the mean minus one standard deviation and the mean plus one standard deviation. And, intuitively, that's just silly as a description of "typical" population: What is a negative population?

These numbers are fail to do their job, which is to represent what the facts in the picture — I can do better than that without even looking at the data (because I know full well that there are *no* countries with negative populations). And so, in these standard deviations I see both the second criterion of well-behaved variables being violated — the standard deviations change sharply between 1975 and 1990 — and I see the fourth criterion beginning to get shaky — the standard deviation does not describe the picture of the data and its values are difficult to interpret.

The medians and the quartiles, of course, give me interpretable numbers — they have to because they always refer to particular cases. That is one reason why medians and quartiles are used, often, in the process of research (although means and standard deviations are more often what is shown in a final report.) But I also note, somewhat uncomfortably, that there is a substantial difference between the "average" I get from one method versus the average I get from the other, between the mean and the median — two very different reports about the middle of the distribution: The median says the average is 6 million people while the mean says the average is 26 million people (for 1975). The median says the average is 6 million people while the mean says the average is 30 million people (for 1990). Now, its not that I can't cope with these numbers and their oddities. Such peculiarities are typical of badly-behaved variables and I can cope with them if I must. But I don't need to.

Now, let's look at the logarithms of these numbers, changing the unit of measure to the logarithm of population, using logarithms base 10.

	Nation	Population in 1,000's		logarithms Base 10	
		1975	1990	1975	1990
1	Tuvalu		9		0.954
2	San Marino		23		1.362
3	Liechtenstein		28		1.447
4	Saint Kits and Nevis		40		1.602
5	Andorra		52		1.716
6	Antigua and Barbuda		64		1.806
7	Aruba		64		1.806
8	Seychelles	60	68	1.778	1.833
9	Kiribati		70		1.845

10	Grenada	100	84	2.000	1.924
11	Dominica		85		1.929
12	Saint Vincent and the Gr	80	113	1.903	2.053
13	Sao Tome and Principe		125		2.097
14	Santa Lucia		150		2.176
15	Vanuatu		165		2.217
16	Western Somoa	160	186	2.204	2.270
17	Maldives	120	218	2.079	2.338
18	Belize		220		2.342
19	Bahamas	200	249	2.301	2.396
20	Barbados	245	254	2.389	2.405
21	Iceland	216	257	2.334	2.410
22	Djibouti		337		2.528
23	Malta	329	353	2.517	2.548
24	Equatorial Guinea	313	369	2.496	2.567
25	Brunei		372		2.571
26	Cape Verde	292	375	2.465	2.574
27	Luxembourg	342	384	2.534	2.584
28	Suriname	422	397	2.625	2.599
29	Comoros	306	460	2.486	2.663
30	Qatar	90	491	1.954	2.691
31	Bahrain	260	520	2.415	2.716
32	Cyprus	673	702	2.828	2.846
33	Fiji	577	738	2.761	2.868
34	Guyana	791	753	2.898	2.877
35	Swaziland	469	837	2.671	2.923
36	Gambia	509	848	2.707	2.928
37	Guinea-Bissau	525	999	2.720	3.000
38	Gabon	521	1,068	2.717	3.029
39	Mauritius	899	1,072	2.954	3.030
40	Botswana	691	1,224	2.839	3.088
41	Trinidad and Tobago	1,009	1,271	3.004	3.104
42	Namibia		1,453		3.162
43	Oman	770	1,481	2.886	3.171
44	Bhutan	1,173	1,566	3.069	3.195
45	Estonia		1,584		3.200
46	Lesotho	1,148	1,755	3.060	3.244
47	Mauritania	1,283	1,935	3.108	3.287
48	Kuwait	1,085	2,124	3.035	3.327
49	Mongolia	1,446	2,187	3.160	3.340
50	Congo	1,345	2,242	3.129	3.351
51	UAR United Arab Emira	220	2,254	2.342	3.353
52	Panama	1,678	2,425	3.225	3.385
53	Jamaica	2,029	2,469	3.307	3.393
54	Liberia	1,708	2,640	3.232	3.422
55	Latvia		2,695		3.431
56	Singapore	2,248	2,721	3.352	3.435
57	Central African Republi	1,790	2,877	3.253	3.459
58	Costa Rica	1,994	3,033	3.300	3.482
59	Uruguay	3,108	3,102	3.492	3.492
60	Albania	2,482	3,273	3.395	3.515
61	Jordan	2,688	3,273	3.429	3.515

62 New Zealand	3,031	3,296	3.482	3.518
63 Lebanon	2,869	3,339	3.458	3.524
64 Armenia		3,357		3.526
65 Ireland	3,131	3,500	3.496	3.544
66 Nicaragua	2,318	3,602	3.365	3.557
67 Turkmenistan		3,658		3.563
68 Togo	2,248	3,674	3.352	3.565
69 Lithuania		3,726		3.571
70 Papua New Guinea	2,716	3,823	3.434	3.582
71 Laos	3,303	4,024	3.519	3.605
72 Sierra Leone	2,983	4,166	3.475	3.620
73 Libya	2,255	4,223	3.353	3.626
74 Norway	4,007	4,253	3.603	3.629
75 Moldova		4,393		3.643
76 Kyrgystan		4,394		3.643
77 Israel	3,417	4,436	3.534	3.647
78 Bosnia Herzegovina		4,517		3.655
79 Paraguay	2,647	4,660	3.423	3.668
80 Benin	3,074	4,674	3.488	3.670
81 Croatia		4,686		3.671
82 Honduras	3,037	4,804	3.482	3.682
83 Finland	4,652	4,977	3.668	3.697
84 Chad	3,947	5,017	3.596	3.700
85 Denmark	5,026	5,131	3.701	3.710
86 El Salvador	4,108	5,310	3.614	3.725
87 Tajikistan		5,342		3.728
88 Georgia		5,479		3.739
89 Burundi	3,765	5,646	3.576	3.752
90 Haiti	4,552	6,142	3.658	3.788
91 Somalia	3,170	6,654	3.501	3.823
92 Switzerland	6,535	6,742	3.815	3.829
93 Bolivia	5,410	6,989	3.733	3.844
94 Kampuchea / Cambodia	8,110	6,991	3.909	3.845
95 Dominican Republic	5,118	7,241	3.709	3.860
96 Guinea	4,416	7,269	3.645	3.861
97 Rwanda	4,233	7,609	3.627	3.881
98 Austria	7,538	7,644	3.877	3.883
99 Senegal	4,418	7,714	3.645	3.887
100 Niger	4,600	7,879	3.663	3.896
101 Tunisia	5,747	8,104	3.759	3.909
102 Mali	5,697	8,142	3.756	3.911
103 Zambia	5,004	8,154	3.699	3.911
104 Angola	6,394	8,449	3.806	3.927
105 Sweden	8,291	8,526	3.919	3.931
106 Bulgaria	8,793	8,934	3.944	3.951
107 Guatemala	6,129	9,038	3.787	3.956
108 Burkina		9,078		3.958
109 Malawi	4,909	9,197	3.691	3.964
110 Yemen		9,746		3.989
111 Serbia		9,883		3.995
112 Belgium	9,846	9,909	3.993	3.996
113 Greece	8,930	10,028	3.951	4.001
114 Belarus		10,257		4.011

115 Portugal	8,762	10,354	3.943	4.015
116 Zimbabwe	6,272	10,394	3.797	4.017
117 Ecuador	7,090	10,507	3.851	4.021
118 Hungary	10,534	10,569	4.023	4.024
119 Cuba	9,481	10,620	3.977	4.026
120 Cameroon	6,433	11,092	3.808	4.045
121 Madagascar	8,020	11,801	3.904	4.072
122 Ivory Coast / Cote d'Ivc	4,885	12,478	3.689	4.096
123 Syria	7,259	12,483	3.861	4.096
124 Chile	10,253	13,083	4.011	4.117
125 Mozambique	9,223	14,539	3.965	4.163
126 Netherlands	13,599	14,936	4.134	4.174
127 Ghana	9,873	15,130	3.994	4.180
128 Afghanistan	19,280	15,564	4.285	4.192
129 Czechoslovakia	14,793	15,683	4.170	4.195
130 Australia	13,809	17,037	4.140	4.231
131 Saudi Arabia	8,966	17,116	3.953	4.233
132 Sri Lanka	13,986	17,198	4.146	4.235
133 Malaysia	12,093	17,556	4.083	4.244
134 Uganda	11,353	18,016	4.055	4.256
135 Iraq	11,067	18,782	4.044	4.274
136 Nepal	12,572	19,146	4.099	4.282
137 Venezuela	12,213	19,698	4.087	4.294
138 Taiwan / Republic of C	16,453	20,435	4.216	4.310
139 Uzbekistan		20,569		4.313
140 Korea North	15,852	21,412	4.200	4.331
141 Peru	15,326	21,906	4.185	4.341
142 Romania	21,178	23,273	4.326	4.367
143 Kenya	13,251	24,342	4.122	4.386
144 Algeria	16,792	25,337	4.225	4.404
145 Morocco	17,504	25,630	4.243	4.409
146 Tanzania	15,388	25,971	4.187	4.414
147 Sudan	18,268	26,245	4.262	4.419
148 Canada	22,801	26,538	4.358	4.424
149 Argentina	25,384	32,291	4.405	4.509
150 Colombia	25,890	33,076	4.413	4.520
151 Zaire	24,450	36,613	4.388	4.564
152 Poland	33,841	37,777	4.529	4.577
153 SPAN	35,433	39,269	4.549	4.594
154 South Africa	24,663	39,539	4.392	4.597
155 Burma	31,240	41,277	4.495	4.616
156 Korea South	34,663	42,792	4.540	4.631
157 Ethiopia	28,134	51,407	4.449	4.711
158 Ukraine		51,711		4.714
159 Egypt	37,543	53,212	4.575	4.726
160 Thailand	42,093	56,002	4.624	4.748
161 France	52,913	56,358	4.724	4.751
162 Iran	32,923	57,003	4.517	4.756
163 Turkey	39,882	57,285	4.601	4.758
164 United Kingdom	56,427	57,366	4.751	4.759
165 Italy	55,023	57,664	4.741	4.761
166 Germany West	61,682	63,232	4.790	4.801
167 Philippines	44,437	64,404	4.648	4.809

168	Vietnam	43,451	66,171	4.638	4.821
169	Germany		79,123		4.898
170	Mexico	59,204	88,010	4.772	4.945
171	Bangladesh		113,930		5.057
172	Pakistan	70,560	114,649	4.849	5.059
173	Nigeria	63,049	118,819	4.800	5.075
174	Japan	111,120	123,567	5.046	5.092
175	Russia		148,254		5.171
176	Brazil	109,730	152,505	5.040	5.183
177	Indonesia	136,044	190,136	5.134	5.279
178	USA	213,925	250,410	5.330	5.399
179	India	613,217	852,667	5.788	5.931
180	China / People's Repu	838,803	1,133,683	5.924	6.054
181	Yemen (Aden)	1,660		3.220	
182	Puerto Rico	2,902		3.463	
183	Hong Kong	4,225		3.626	
184	Upper Volta	6,032		3.780	
185	Yemen (Sana)	6,668		3.824	
186	Germany East	17,127		4.234	
187	Vietnam South	19,650		4.293	
188	Yugoslavia	21,322		4.329	
189	Vietnam North	23,800		4.377	
190	Bangladesh	73,746		4.868	
191	USSR	255,038		5.407	
192	Burkina				
193	Soviet Union frmr				

Beginning again, the first thing you “see” using logs is that the numbers are unfamiliar. That’s not good, it leads to error, so I’ve introduced a few bench marks by using logs base 10. Using logs base 10, a “2” in logs, corresponds to 100 without logs. So, St. Vincent with approximately 100 (approximately one hundred thousand people) will have a log, base 10, of approximately 2. Using logs base 10, a “3” in logs corresponds to 1,000 without logs. So Guinea-Bissau with approximately 1,000 (approximately one million people) will have a log, base 10, of approximately 3. Belgium with approximately 10,000 (approximately ten million people) will have a log, base 10, of approximately 4. And, while you will become accustomed to these numbers, there is no harm done by keeping the original values in the table, for backup.

Now for the picture, the shape of the stem and leaf. Preparing to select boundaries for the stems, I check the range, finding a range between .954 and 6.05. Using convenient boundaries to get about ten stems, and just checking the counts I would get for these stems, I get Figure _. I will not actually construct the stem and leaf because with the rank ordering in Figure _, including the names, and the shape shown in Figure _, I have what I need.

```
0.5-    .999    |  |                                                        1 country
1.0 -   1.499   |  ||                                                       2 countries
1.5 -   1.999   |  ||||||||                                                 8countries
2.0 -   2.499   |  ||||||||||                                               10 countries
2.5 -   2.999   |  |||||||||||||||                                          15 countries
3.0 -   3.499   |  |||||||||||||||||||||||                                  23 countries
3.5 -   3.999   |  |||||||||||||||||||||||||||||||||||||||||||||||||||||    53 countries
4.0 -   4.499   |  ||||||||||||||||||||||||||||||||||||                     36 countries
4.5 -   4.999   |  ||||||||||||||||||||||                                   22 countries
5.0 -   5.499   |  ||||||||                                                 8 countries
5.5 -   5.999   |  |                                                        1 country
6.0 -   +       |  |                                                        1 country
```

That’s good — decidedly closer to symmetry than the original and suggesting that log population may be a well-behaved (or relatively well-behaved) variable. Let’s find the numbers that are supposed to describe the picture.

	1975	1990	
Mean	3.724	3.688	**Verify**
Standard Deviation	.787	.911	**Verify**
(Excel function stdevp)			
Median	3.759	3.8055	**Verify**
Low Quartile	3.3035	3.222	**Verify**
High Quartile	4.2295	4.278	**Verify**
Quartile Spread	.926	1.056	

Looking at the averages:

This is certainly different. Let's take it apart. The mean is approximately 3.7 to 3.8 and so is the median. Whichever average I choose, I get approximately the same result. That's another sign of symmetry (Symmetry will place the both the median and the mean in the middle, near the peak. So this is another suggestion that the variable is well-behaved.) And keeping myself grounded in the units for which I have intuition, 3.7 is the log, base 10, of a number slightly less than 10,000 (with 4 zeroes), implying approximately 10 million people as the central value for the distribution of populations — approximately the size of Denmark, the Dominican Republic, or Guatemala and probably a bit of a jolt to the intuition of an American living in a nation of 250 million).

Between 1975 and 1990, the mean has actually decreased a little, that's a little unsettling. Is this some weird result of using a weird unit of measure, the logs? No, but let's check: My intuition expects an increase of about two percent per annum, for world population, approximately 30% over the fifteen years between 1975 and 1990. (I got 30% for fifteen years by multiplying 2% per year, my personal expectation, times 15 years. That's a crude approximation because it ignores compound interest but then I pulled 2% out of my head anyway. For now, for the sake of a *quick* look and a quick think about the data, I can tolerate the crudeness of these approximations.)

But the mean population (in logs) hasn't increased at all and the median population (in logs) has increased by .0465. What is .0465? That is the difference between the logs so it corresponds to a ratio of populations, 10.0465 = 1.11. So it says that the 1990 figure is 1.11 times the 1975 figure, an increase of 11%, not 30%. So both figures for the 1975 to 1990 change are too low — compared to what I am expecting. Again, is this just a peculiar punishment for using peculiar numbers, the logs? No, it can't be because the median country is the same country, regardless of the choice between population and log population. So there is something real here.

Unless — maybe my figure of 2% per annum for world population is wrong. I don't know how that number got into my head. And since my intuition is not matching the facts, I'd better check. And, I have the data: Adding up the 1975 populations, I get about 4 billion people. Adding up the 1990 populations, I get 5.3 billion. That's an increase of about 33% in *total* population — right on target for my intuition. So I still have to explain the difference between the increase for *total* population and the increase for *average* population. Ah — this begins to sound like a problem of units: For total population the unit is the world. For average population the unit is the nation. So, I've got a clue but I've still got something to worry about as I continue.

Looking at the variations:

What have I got for the variation? First, what am I expecting? I'm expecting or, to be more precise, I am hoping that the variation is a nice reasonable number — unlike the variation that including a negative range of populations (for population without logs). Taking the measures of variation one at a time (without comparing them), this part looks good. Using the standard deviations, the central range of the distribution is the range between the mean minus one standard deviation and the mean plus one standard deviation — for 1990 that's between 2.777 and 4.599. Unlike the measures of variation on population, which was comparatively poorly behaved, these numbers match the picture, including a range from about one and a half stems less than the stem with the largest number of leaves to about one and one half stems greater than the stem with the largest number of leaves. Bringing my intuition along, what is .911? It is the log of 8.15, (10^.911 = 8.15). That tells me that the central range of the distribution lies between the center divided by 8.15 and the center multiplied by 8.15.

To make that a little easier, in plain English, I can introduce the term "geometric mean": The geometric mean is the anti-log of the mean of the logs, i.e., the anti log of 3.688 in this case. So, I can build this information into a sentence by saying that "the central range of populations lies within a factor of 8 on either side of the geometric mean of the populations."

The quartile spreads, used as alternative way to specify the central range of the distribution, pick out a narrower definition of the center, between 3.222 and 4.278. I would put that in words by using anti-logs and saying "the middle fifty percent of the distribution marks a range between 1.7 million and 19 million around a median of 6.4 million people.

What have I got for the comparison between variations? First, again, what am I expecting? I'm expecting, or hoping, that the variation is constant, more evidence of a well-behaved variable. Checking the facts, no. Both the standard deviations and the quartile spreads increase, 1975 to 1990. The standard deviation increases by .125. The quartile spread increases by .130. Again, how large are these logarithms? Checking, .125 is the log of 1.33, (10^.125 = 1.33). .130 is the log of 1.35. Those ratios seem like large increases, thirty-three to thirty-five percent. I don't like that — not well-behaved. It may be that these are the facts and there is nothing to do but report them. But that's just deferring the thinking — somebody, probably me, is going to have to figure out what's going on here. One strong cue I have is my knowledge that the computation of quartiles doesn't even use the data beyond the quartiles ,while the computation for the standard deviation, uses all of the data. That suggests I keep my eye on the small end or the large end of the distribution of nations. I'll put this on the stack of questions I have to worry about as I continue.

Looking at the Limits:

Checking the upper and lower ends of the 1990 distribution: The difference between results based on the given unit of measure, population, as compared to the well-behaved unit of measure, log population, will differ most at the low end: Where the mean minus one standard deviation extended to negative populations, using people; the mean minus one standard deviation is still a credible number, using logs.

Doing the computations: Using two criteria, first using the criterion that builds on the standard deviation. Using straight population, no logs, the mean minus three standard deviations and the mean plus three standard deviations set the bounds of reasonable data at –305,141, that's negative 305 million, and +364,555, that's plus 365 million. Aside from the fact that there is no country with a negative population, whatever that means, there are three countries, the United States, India, and China, with unusually large populations. But, such measures don't make sense for a variable, population, that is not well-behaved.

Applying the same computations using logs as the unit of measure, three standard deviations sets the bounds of reasonable data at 1.855 and 5.521, corresponding to populations below 71.6 (below 72 thousand) and above 331,894.5 (above 339 million), marking one country, Seychelles, as atypically small and two countries, India and China, as atypically.

Using the criteria based on quartiles, suggested by Tukey, the inner fences are at 2.2215 and 5.3895, corresponding to 166.5 (167 thousand) and 245,188 (two hundred and forty five million). And the outer fences are at .6375 and 6.9735, corresponding to 4.34 (four thousand) and 9,408,058 (9.4 billion). Using the inner fences, they classify 15 countries as very small and three as very large. Using the outer fences, no countries are "beyond the fence".

Summing it up and thinking:

So now, what have I got: The mean has become smaller while the median has become larger. This too may have something to do with using only the middle values, for the median, while using all the values for the mean. And now I also remember that there was a substantial difference between the list of nations for 1975 and the list for 1990. Looking at that list again, in rank order, I see one very big country has disappeared, replaced by ?? smaller ones. So I bet that the anomalies I need to deal with, one average increasing while the other decreases, variations that get larger when, for I was looking for a constant, that these anomalies can depend on something that occurred among the extreme values, suggesting that this is the numerical indicator of the breakup of the Soviet Union.

I'll check. Removing the Soviet Union from the 1975 data produces

	1975	1990	
Mean	3.713	3.683	**Verify ??**
Standard Deviation (Excel function stdevp)	.777	.936	**Verify**
Median	3.7575	3.8445	**Verify**
Low Quartile	3.3000	3.1665	**Verify**
High Quartile	4.225	4.289	**Verify**
Quartile Spread	.925	1.056	

So much for that explanation, it doesn't work. — I've removed the USSR from the 1975 data and Estonia, Latvia, Armenia, Turkmenistan, Lithuania, Moldova, Kurqustan, Tajikistan, Georgia, Belarus, Uzbekistan, and the Ukraine from the 1990 data — but the effects I hoped to explain have persisted.

I'd better be more extreme: I'll use only those countries presenting data for both time periods: At 144 countries, this is a restricted subset, but it gives me a subset whose change, 1975 to 1990, depend more (though not exclusively) on population growth than political change (though not exclusively) — excluded other nations for the purpose of checking whether the curious changes of values are attributable to fission and fusion.

	1975	1990	
Mean	3.693	3.839	**Verify**
Standard Deviation (Excel function stdevp)	.790	.777	**Verify**

Median	3.721	3.892	**Verify**
Low Quartile	3.2425	3.4075	**Verify**
High Quartile	4.1935	4.336	**Verify**
Quartile Spread	.9510	.9285	

All right — somewhere between the full data set, which is different for the two dates, and the restricted data set, which is more similar for the two dates (though less complete for either date) some of the directions of change have reversed. With full data the mean decreases, with restricted data, the mean increases. With full data the standard deviations increase; with restricted data the standard deviations decrease. With the full data as well as the restricted data the medians increase. With full data the quartile spread increases; with restricted data the quartile spread decreases.

I'm now a little more at ease about the logarithm. It was supposed to give me a well-behaved variable. And it did gain symmetry and reasonable spreads. But it bothered me when my two measures of spread, the standard deviations and the quartile spreads of the logs increased. Now they decrease. That tells me I'm close: The size of the changes in the variation are within a range that can be influenced by changes in the set of countries. So the variations do not necessarily indicate that the measure is poorly-behaved. Phrasing that with a double negative: I have no clear evidence that log population is not homeoscedastic. (Try these tests with the original numbers — in people as the unit the variable. The heteroscedasticity there **are/ is /should be [check]** much larger than might be explained by heteroscedastic with or without the adjustments in the set of nations.)

The Write Up

And now, wanting you to know how hard I've worked at this, but admitting, that you really don't care — except for the results: The write up:

The Size of Nations

What is the size of a country? United Nations data for 1990 show that the median population of countries is about 5 million, matching the population of Somalia and Haiti. Those of us who live in the United States naturally think of ourselves as "normal" and countries such as Somalia and Haiti as quite small, but in fact the middle fifty percent of national populations runs between 1.7 million, e.g., Estonia, and 19 million, Iraq. At the extremes, China and India stand out, including between themselves, 2 billion people, about half of the world's population while, at the other end a few nations count populations of approximately 100,000 or less.

Curiously, even while the population of the world has increased by about 35%, approximately 2% per year, the population of the average nations shows no clear trend. Even the direction of change, up or down, depending on the precise measure that is used to compute the average . (The size of the median population has increased from 5.7 million to 6.4 million, a 12% increase.) In effect, as the total population has increased, the division of people into states has increased the fragmentation so that the 1990 average is close to the 1975 average and smaller than the 1975 average as a fraction of the total population of the world. This fragmentation appears to be the net result of changes in definitions and borders, including the breakup of the Soviet Union and Yugoslavia, the fusion of the Germanys, Vietnams, and Adens.

Who, What, Where …

I've chosen the median because I can identify it with a country. That makes it easy to communicate. It also avoids the need to discuss the units because the median is the median (the middle is the middle) whether I use the logs or the original numbers.

I'm using bench marks to keep my reader oriented, using the size of the U.S. the names of well known countries, and rounded numbers.

These were convenient cutting points, using either the ratio of the size of a country to the size of the next smallest country, 3.4 for India as compared to China, 1.3 for China as compared to India, or easily remembered values, like 100,000.

I've chosen to make a virtue out of ambiguity with respect to the direction of change, acknowledging different results from different indicators.

I'm giving up on the distinction between nation and state — important professionally but not necessarily to my audience (depending on the audience)/

The Unit of Analysis: Facts About What?

O.K., if you didn't already know how, then now you know how to compute a median and a mean. Those are the centers. Corresponding to each of these concepts of the center there is an associated concept of the variation: Corresponding to the median, the hinges provide a way to specify, verbally, where the central half (not just the center) of the data lies. In words, using the median and the hinges: "The middle 50% lie between ___ (lower hinge) and ___ (the upper hinge). Corresponding to the mean, the standard deviation estimates the mean deviation: "The mean is ___ with a standard deviation of ___." The verbal summary provides the mean and the standard deviation from which the reader of the summary is supposed to construct a mental image a bell-shaped distribution with the central peak at the mean and the center of the distribution lying between values which are one standard deviation below the mean and one standard deviation above the mean.

Corresponding to each of these concepts of variation there is also a concept of *too much* variation: Corresponding to the median and the quartiles, too much variation is marked by the fences: The inner and outer fences mark the limits of routine variation. Value beyond the fences are sufficiently unusual to be suspicious — not just different from the central values of the data, but different in kind — or, at the least, that is a possibility to be investigated. Corresponding to the mean and the standard deviation, too much variation is marked by the value of the mean plus or minus two standard deviations or three standard deviations (or, sometimes, more precisely plus or minus 1.96 and 2.81 standard deviations). In either system "too much variation" designates variation so large and unexpected that the analyst may either leave it out entirely, applying the name "outlier", or do just the opposite by focusing in on these special cases as extreme examples of the general principles at work in the data (Gerber's High Protein).

In addition to learning the concepts for the center, for the variation, and for too much variation, you have taken steps toward sorting all of this out computationally, and you've begun to get Excel and the Mac to do your bidding.

With these "mechanics" under control, it is time to delve in to the art of using these things, time to think. The first question is "Facts about what?" A datum is an attribute of something: A number of grams of protein is an attribute of a cereal. A cause of death is an attribute of a person. A literacy rate is an attribute of a country. I want to draw your attention to that thing: What is this thing to which the numbers are attached or, in the language of the trade "What is the correct unit of analysis?"

The routine answer to that question is usually easy enough to answer. The deeper answer is more interesting: I asked "What *is the correct* unit of analysis?" There are choices. And there is no compelling reason why the data analyst should automatically accept the unit of measure that was convenient to the person who organized the data. That is up to us — analyst's choice. And the choice may make a difference. For example, beginning with the breakfast cereal, the original data show grams of protein as an attribute of a commonly used portion of breakfast cereal . But grams of protein could be recomputed as an attribute 100 grams of breakfast cereal — changing the unit of analysis by standardizing the data to a common weight of cereal. Analyst's choice.

To emphasize the importance of the unit of analysis and to encourage active choice of the unit of analysis on the part of the data analyst, I am going to take you through five sets of mental gymnastics. The job of these gymnastics is to create doubt, doubt with respect to passive acceptance of the data as given, and then to improve the focus of the analysis by asking "What is the 'correct' unit of analysis?"

On the Average I:
Physicians per Capita

Data gymnastics, exercise #1: In Figure 1 (attached Excel file 1 Phy/Cap Rdcd Cols) you have data for 137 countries: Column 1 names the country. Column 2 gives you an estimate of the number of physicians in that country. Column 3 gives you the estimate of the population for that country. Question: What is the average rate of physicians per capita?

As a problem in arithmetic, that's easy: For each nation I construct the ratio of doctors to people. That gives me a new number describing each nation, computed and shown in columns 4 and 5. So, for the United States (in 1975), row 130, the data show 348-thousand doctors and 213-million people, which works out to .001629 doctors per person, 1.6 doctors per thousand people. I repeat that computation for each row, getting a number that describes each nation. And then I compute the average: The average of these national statistics is .000681 physicians per capita. That's 0.7 physicians per thousand, approximately two thirds of a doctor for each one thousand people.

That's it — or is it? It's good to do things, even easy things, two or more ways, just to be sure that everything" is right, just to be sure that everything is clear. So looking at these data for a second time, surely there's an easier way than the one I used above: If I want the number of physicians per person in the world (that part of the world for which I have data), then why not just add up the number of doctors, add up the number of people, and then divide the number of doctors by the number of people? Why not?

Adding-up the number of doctors: The sum is approximately three million doctors (reported as 2,622,088 doctors at the bottom of Column 2). Adding-up the number of people: The sum is approximately three billion, (reported as 3,028,196,000 noted at the bottom of Column 3). So, dividing the number of doctors by the number of people, the answer is

.000866 doctors per capita, which is approximately .001 doctors per capita or, approximately 0.9 doctors per thousand people.

And, now, as you see, I have a problem: The first time through, I got 0.7 doctors per thousand people. The second time through I got 0.9 doctors per thousand people. And, of course, 0.9 just is not the same answer as 0.7. If you accepted what I did, and found both procedures straight forward (and numerically correct), then why are the two answers different?

Should I worry about such a thing? I could get out by talking quickly, flashing a few extra digits and then concluding that I have two estimates showing between one-half and one doctor per thousand people, across the world in 1975. Or I could dismiss the whole thing, saying that the two numbers are almost the same. But that's just whistling in the dark and hoping no one will notice that I'm in trouble. If you accepted both arguments and if both arguments were correct, then the numbers should be exactly the same. Something is wrong.

I could take another out: Surely all of these data are rough approximations at best. Possibly some of them are as much bravado as fact establishing a nation's standing in the pecking order of nations — my country is better than yours , when the United Nations compares one nation to another. And the data are suspect: If you do your homework, beginning with one variable checks for each of the two variables, and continue with your one variable check on the column of numbers for doctors per capita, then you will have some serious questions about this stuff. But that's just another verbal dodge: confusing my reader by arguing that fuzzy data allow fuzzy conclusions, even for something as simple as the average. That's a dodge: Whether or not I believe the detailed numbers, my problem is that I have what appear to be two different estimates of doctors per person both based on the same numbers — that's not acceptable because it shows that something is wrong — it is a loose thread in my web of credibility and I've got to fix it.

Well, there's another possibility here that may rescue me from confusion: There is an awful lot of estimation going on in these numbers, some of them are large numbers, some of them are small — and that's the kind of thing that leads computers into rounding error. That's a possibility to be considered. It's not the real problem here but it leads to an important suggestion: We've got two confusing things going on here: We've got data that we're trying to understand and, as it turns out, we've got a method that we've got to work on before the method itself is understand. And, for the moment, both of them are confusing. So, let's simplify the problem, simplify the data, in order to focus on the complexity of the method.

Country	Doctors	Total Population '75		Doctors Per Person	Doctors Per 1,000 People	
Bahamas	161	200,000		0.0008005	0.805	
Barbados	166	245,000		0.000677551	0.678	
Canada	39,104	22,801,000		0.001715012	1.715	
Costa Rica	1,292	1,994,000		0.000647944	0.648	
Cuba	8,201	9,481,000		0.000864993	0.865	
El Salvador	1,117	4,108,000		0.000271908	0.272	
Grenada	25	100,000		0.00025	0.250	
Guatemala	1,207	6,129,000		0.000196933	0.197	
Guyana	237	791,000		0.000299621	0.300	
Haiti	396	4,552,000		8.69947E-05	0.087	
Honduras	920	3,037,000		0.000302931	0.303	
Jamaica	570	2,029,000		0.000280927	0.281	
Mexico	31,556	59,204,000		0.000533005	0.533	
Nicaragua	1,400	2,318,000		0.000603969	0.604	
Panama	1,404	1,678,000		0.00083671	0.837	
Puerto Rico	3,479	2,902,000		0.001198828	1.199	
Trinidad and Tobago	550	1,009,000		0.000545094	0.545	
USA	348,484	213,925,000		0.001629001	1.629	
Sums (for Doctors & Population)	440,269	336,503,000	Averages:		0.653	

Doctors/Person =		***0.001308366***				
Doctors/Thousand Persons =		***1.308***		***Compare as Doctors per 1,000 People***		

Figure _ (Excel file Phy/Cap Rdcd Cols 18 Rows) shows the same kind of data, restricted to North American and Caribbean nations. The problem is still there: The two different "averages" remain unequal, shown in the same locations as Figure 1. But there's no reason to stop here: For the moment, I don't need a set of nations that represents anything, my immediate problem is to understand the method — then and only then will I try to use it. So simplifying, here is a "sample" of the U.S. and Mexico. Let's take a look at the data for these two.

Country	Doctors '75	Population '75	Doctors Per Person	Doctors Per 1,000 People
Mexico	31,556	59,204,000	0.000533	0.533
USA	348,484	213,925,000	0.001629	1.629
			Averages:	
Sums (for Doctors & Population)	380,041	273,129,000	0.001081	1.081 925 people per doctor
Doctors/Person =		0.00139143		
Doctors/Thousand Persons =		1.391 719 people per doctor		

Derived from: World Handbook of Political and Social Indicators, Third Edition, by Char Taylor and David A. Jodice, Yale University Press, New Haven and London, 1983

The first point is that the inequality is still there: With only two countries left in the table the two different estimates of the "average" are still different, 1.081 versus 1.391. And there is nothing much left that can be dismissed as detail or "complication". Whatever is going on, whatever it is that makes these two computations represent different things not just different computations. Do I care about this 20 to 25% discrepancy. You bet: Unless I understand what's going on here, all I know is that I've got a problem and I don't even know how large it is, or how large it would be in other data — the discrepancy tells me I don't know what I'm doing — and I'm responsible for that. And if this were not an exercise but the beginning of a research project, then work would stop right here, or should stop, until the principles are understood: If my research required me to estimate how much the ratio of doctors to people had changed in twenty-five years, if my research needed to ask whether socialized medical systems were different from others, if my inquiry were to ask how physicians per capita was related to nations wealth and, perhaps, how physicians per capita was inversely related to other government expenditures — whatever it is that I need to know for research or policy — I'm surely not ready for subtle comparisons among nations when I can't even zero-in on the first estimate for the first set of data. I need an answer.

The answer, or at least the route to an answer, lies in looking very closely at the unit of analysis: What is it that is described by each of the numbers? So let's add some units and then carry the units into the arithmetic: Here are the "data", the basic four numbers with labels.

31,556 doctors	59,204,000 people
348,484 doctors	213,925,000 people

Country	Doctors	Population	Doctors	Doctors
			per	per 1,000
			Person	People
Mexico	31,556	59,204,000	0.000533	0.533
USA	348,484	213,925,000	0.001629	1.629

When I compute the first ratio, .533 doctors per thousand people, that ratio describes the unit Mexico: It had half a doctor per thousand people. When I compute the second ratio, 1.629 doctors per thousand, the ratio describes the unit United States: At one and one half doctors per thousand people, the United States has approximately three times as many doctors per capita as Mexico. Each of the two numbers measures the attribute of a country. And the average with respect to these two countries is the average of doctors per person *per country*. 1.1 doctors per thousand people is the single number that comes closest to describing these two data points

By contrast, the second computation adds up the numbers of doctors to estimate the number of doctors in the world (in the two country example, it is the “world” comprised of these two countries). It adds up the number of people to estimate the number of people in the world. And then, dividing the number of doctors in the world by the number by the number of people in the world it describes the average number of doctor per person in the world at large — ignoring countries.

At risk of going past clarity to something that is “painfully” clear, the more precise description of these data requires, and uses, two levels of aggregation, each of which corresponds to a unit of analysis. At the first level of aggregation the unit of analysis is the person. It shows the number of doctors *per person*.

One of the two descriptions ascends to a second level of aggregation, the country. Here doctors per person is an attribute of the country which is the unit of analysis. In Mexico, there are .0005 doctors per person, 0.5 doctors per thousand. In the United States, there are 1.6 doctors per thousand. And the average of the two numbers is the doctors per person of an average country.

By contrast, the other computation continues to compute doctors per person assembling one datum (not really an average) that describes the world. It is equally correct to say that in this second analysis the unit of analysis is the person. In the world (the “world” of the United States and Mexico) there are about 1.4 doctors per thousand people.

Aside on homework:

When you try to explain the difference between these two averages there is at least one very sophisticated way of explaining it which is both sophisticated and wrong.

Here's the wrong way:

> "The average that adds up the two numbers and divides by two treats both countries equally. But they are not equal. The United States has a larger population than Mexico. So the average computed by adding two numbers and divides by two gives less weight to the average American than it does to the average Mexican. By contrast, summing the populations, summing the doctors and then dividing one sum by the other gives all people equal weight."

That explanation is wrong. It discusses the weight given to each person in one computation versus the weight given to each person in the other population. And therefore, subtly, it implies that the person is the correct unit of analysis.

That explanation is wrong because it misses the point. Sure enough, we can use weights to convert the data into results that describe one unit of analysis or another unit of analysis. But the important point *is* that unit of analysis. The country as a unit of analysis has an economy, a health care system which includes medical education, an economy of medical care, and a distribution system: The average system delivers 1.1 doctors per thousand people. By contrast, the person as a unit of analysis, shorn of politics, shorn of the wealth of the country in which a person lives, experiences 0.9 doctors per thousand. That's the experience of the average person.

So, back to the data with which I began. What can I say about doctor per person:

UNESCO data for 1975 estimate 2.6 million doctors and 3.0 billion people in the world, leaving a world wide average of .866 doctors per thousand people, less than one doctor per thousand people. But nations are unequal in size and unequal in their ability to deliver these services, so that the average nation-based health system achieves only .681 doctors per thousand — which implies that some of the larger nations have relatively inferior delivery of services, inferior as compared to the average nation.

In this regard it should be noted that data for China are missing from these numbers. With 20% of the world's population excluded from the data, the estimate of the resources available to the average person must be treated with caution. However as China is but one nation among many, the description of estimate of results achieved by the average health care system is not seriously affected by the absence of data for China.

Homework: You will receive an Excel sheet describing, one, the population, and, two, the Gross National Products of nations. First, compute the gross national product *per capita* for each nation and compute the average of the gross national products. Second, add-up the gross national products and add up the populations and compute the product per person. That's the easy part. Now: Write a short essay clearly describing the two results and what each of them means.

I give you a target audience for your essay: I like to try to write for my ten year old niece. What is a ten year old? A ten year old has all the operating mental equipment of an adult, but none of the experience, and a very short attention span. The ten year old is not easily impressed or intimidated — efforts to do so usually just lose the attention of such an audience: You have to get to the point. Your "story" has to connect the numbers and it has to make a point. And you have to make it short.

And also: Be very careful with the units. I've given them to you as they come from the data base I am using: thousands of people and millions of dollars, not people and dollars. Above, I hid my clean up of the data base, not showing you the numbers until after I had put them in terms of people and doctors. I don't mind if you work with data tabulated in thousands and millions, but be very careful about the units on your results: When you report dollars per person, in either of the two ways you will use, make sure you really have dollars per person.

Country	Doctors	Total Population '75	Doctors/Person	Doctors Per 1,000 People
Afghanistan	656	19,280,000	3.40249E-0	0.034
Angola	384	6,394,000	6.00563E-0	0.060
Argentina	48,687	25,384,000	0.00191801	1.918
Austria	15,702	7,538,000	0.00208304	2.083
Bahamas	161	200,000	0.00080	0.805
Bahrain	177	260,000	0.00068076	0.681
Bangladesh	5,088	73,746,000	6.89936E-0	0.069
Barbados	166	245,000	0.00067755	0.678
Belgium	18,510	9,846,000	0.00187995	1.880
Benin	95	3,074,000	3.09044E-0	0.031
Bolivia	2,581	5,410,000	0.00047707	0.477
Botswana	63	691,000	9.11722E-0	0.091
Brazil	62,656	109,730,000	0.00057100	0.571
Bulgaria	18,773	8,793,000	0.00213499	2.135
Burma	5,561	31,240,000	0.00017800	0.178
Burundi	83	3,765,000	2.20452E-0	0.022
Cameroon	354	6,433,000	5.50288E-0	0.055
Canada	39,104	22,801,000	0.00171501	1.715
Central African Republic	97	1,790,000	5.41899E-0	0.054
Chad	83	3,947,000	2.10286E-0	0.021
Chile	4,419	10,253,000	0.00043099	0.431
Colombia	12,997	25,890,000	0.00050200	0.502
Comoros	21	306,000	6.86275E-0	0.069
Congo	213	1,345,000	0.00015836	0.158
Costa Rica	1,292	1,994,000	0.00064794	0.648
Cuba	8,201	9,481,000	0.00086499	0.865
Cyprus	547	673,000	0.00081277	0.813
Czechoslovakia	35,385	14,793,000	0.0023920	2.392
Denmark	9,896	5,026,000	0.00196896	1.969
Dominican Republic	2,375	5,118,000	0.00046404	0.464
Ecuador	3,517	7,090,000	0.00049605	0.496
Egypt	8,034	37,543,000	0.00021399	0.214
El Salvador	1,117	4,108,000	0.00027190	0.272
Equatorial Guinea	5	313,000	1.59744E-0	0.016
Ethiopia	338	28,134,000	1.20139E-0	0.012
Finland	6,699	4,652,000	0.00144002	1.440
France	77,888	52,913,000	0.00147200	1.472
Gabon	96	521,000	0.00018426	0.184
Germany East	31,308	17,127,000	0.00182799	1.828
Germany West	122,069	61,682,000	0.00197900	1.979
Ghana	938	9,873,000	9.50066E-0	0.095
Greece	18,423	8,930,000	0.00206304	2.063
Grenada	25	100,000	0.0002	0.250
Guatemala	1,207	6,129,000	0.00019693	0.197
Guinea	278	4,416,000	6.29529E-0	0.063
Guyana	237	791,000	0.00029962	0.300
Haiti	396	4,552,000	8.69947E-0	0.087
Honduras	920	3,037,000	0.00030293	0.303

Hong Kong	2,881	4,225,000	0.00068189	0.682
Hungary	21,131	10,534,000	0.00200598	2.006
Iceland	372	216,000	0.00172222	1.722
India	145,946	613,217,000	0.00023800	0.238
Indonesia	8,299	136,044,000	6.10023E-0	0.061
Iran	11,358	32,923,000	0.00034498	0.345
Iraq	4,504	11,067,000	0.00040697	0.407
Ireland	3,773	3,131,000	0.00120504	1.205
Israel	9,144	3,417,000	0.00267603	2.676
Italy	114,228	55,023,000	0.00207600	2.076
Ivory Coast / Cote d'Ivoire	322	4,885,000	6.59161E-0	0.066
Jamaica	570	2,029,000	0.00028092	0.281
Japan	133,344	111,120,000	0.0012	1.200
Jordan	745	2,688,000	0.00027715	0.277
Kenya	1,246	13,251,000	9.40306E-0	0.094
Korea South	17,851	34,663,000	0.00051498	0.515
Kuwait	1,089	1,085,000	0.00100368	1.004
Laos	155	3,303,000	4.69270E-0	0.047
Lebanon	2,301	2,869,000	0.00080202	0.802
Lesotho	49	1,148,000	4.26829E-0	0.043
Liberia	142	1,708,000	8.31382E-0	0.083
Libya	2,586	2,255,000	0.00114678	1.147
Luxembourg	368	342,000	0.00107602	1.076
Madagascar	754	8,020,000	9.40150E-0	0.094
Malawi	103	4,909,000	2.09819E-0	0.021
Malaysia	2,007	12,093,000	0.00016596	0.166
Maldives	9	120,000	0.000075	0.075
Mali	142	5,697,000	2.49254E-0	0.025
Malta	382	329,000	0.00116109	1.161
Mauritania	87	1,283,000	6.78098E-0	0.068
Mauritius	346	899,000	0.00038487	0.385
Mexico	31,556	59,204,000	0.00053300	0.533
Mongolia	2,604	1,446,000	0.0018008	1.801
Morocco	1,243	17,504,000	7.10123E-0	0.071
Mozambique	507	9,223,000	5.49713E-0	0.055
Nepal	339	12,572,000	2.69647E-0	0.027
Netherlands	21,826	13,599,000	0.00160497	1.605
New Zealand	4,110	3,031,000	0.00135598	1.356
Nicaragua	1,400	2,318,000	0.00060396	0.604
Niger	83	4,600,000	1.80435E-0	0.018
Nigeria	4,224	63,049,000	6.69955E-0	0.067
Norway	6,884	4,007,000	0.00171799	1.718
Oman	153	770,000	0.00019870	0.199
Pakistan	17,922	70,560,000	0.00025399	0.254
Panama	1,404	1,678,000	0.0008367	0.837
Paraguay	2,229	2,647,000	0.00084208	0.842
Peru	10,514	15,326,000	0.00068602	0.686
Philippines	13,464	44,437,000	0.00030299	0.303
Poland	58,240	33,841,000	0.00172098	1.721
Portugal	11,101	8,762,000	0.00126694	1.267
Puerto Rico	3,479	2,902,000	0.00119882	1.199
Qatar	96	90,000	0.00106666	1.067
Romania	28,548	21,178,000	0.00134800	1.348

Rwanda	106	4,233,000	2.50413E-0	0.025
Sao Tome and Principe	12	80,000	0.0001	0.150
Saudi Arabia	3,613	8,966,000	0.00040296	0.403
Senegal	305	4,418,000	6.90358E-0	0.069
Seychelles	21	60,000	0.0003	0.350
Singapore	106	2,248,000	4.71530E-0	0.047
Somalia	193	3,170,000	6.08833E-0	0.061
South Africa	12,060	24,663,000	0.00048899	0.489
SPAN	54,992	35,433,000	0.00155	1.552
Sri Lanka	3,245	13,986,000	0.00023201	0.232
Sudan	1,407	18,268,000	7.70199E-0	0.077
Suriname	202	422,000	0.00047867	0.479
Swaziland	65	469,000	0.00013859	0.139
Sweden	14,045	8,291,000	0.00169400	1.694
Switzerland	11,469	6,535,000	0.00175501	1.755
Syria	2,403	7,259,000	0.00033103	0.331
Tanzania	846	15,388,000	5.49779E-0	0.055
Thailand	5,009	42,093,000	0.00011899	0.119
Togo	1,623	2,248,000	0.00072197	0.722
Trinidad and Tobago	550	1,009,000	0.00054509	0.545
Tunisia	1,213	5,747,000	0.00021106	0.211
Turkey	21,696	39,882,000	0.00054400	0.544
UAR United Arab Emirates	681	220,000	0.00309545	3.095
Uganda	431	11,353,000	3.79635E-0	0.038
United Kingdom	75,612	56,427,000	0.00133999	1.340
Upper Volta	109	6,032,000	1.80703E-0	0.018
USA	348,484	213,925,000	0.00162900	1.629
USSR	733,744	255,038,000	0.00287699	2.877
Venezuela	13,105	12,213,000	0.00107303	1.073
Vietnam South	9,000	19,650,000	0.00045801	0.458
Western Somoa	55	160,000	0.0003437	0.344
Yemen (Sana)	367	6,668,000	5.50390E-0	0.055
Yugoslavia	27,143	21,322,000	0.00127300	1.273
Zaire	807	24,450,000	3.30061E-0	0.033
Zambia	470	5,004,000	9.39249E-0	0.094
Zimbabwe	916	6,272,000	0.00014604	0.146

	Sum	Sum		Aver-age
Sums (for Doctors Population)	2,622,08	3,028,196,000		0.681
2.6 million Doctors/ 3.0 billion Persons=		0.000865891		
Doctors/ Persons in Thousands=		0.866	Compare as Doctors per 1,000 People	

Country	Doctors	Total Population '75	Doctors/Person	Doctors Per 1,000 People
Bahamas	161	200,000	0.000805	0.805
Barbados	166	245,000	0.000677551	0.678
Canada	39,104	22,801,000	0.001715012	1.715
Costa Rica	1,292	1,994,000	0.000647944	0.648
Cuba	8,201	9,481,000	0.000864993	0.865
El Salvador	1,117	4,108,000	0.000271908	0.272
Grenada	25	100,000	0.00025	0.250
Guatemala	1,207	6,129,000	0.000196933	0.197
Guyana	237	791,000	0.000299621	0.300
Haiti	396	4,552,000	8.69947E-05	0.087
Honduras	920	3,037,000	0.000302931	0.303
Jamaica	570	2,029,000	0.000280927	0.281
Mexico	31,556	59,204,000	0.000533005	0.533
Nicaragua	1,400	2,318,000	0.000603969	0.604
Panama	1,404	1,678,000	0.00083671	0.837
Puerto Rico	3,479	2,902,000	0.001198828	1.199
Trinidad and Tobago	550	1,009,000	0.000545094	0.545
USA	348,484	213,925,000	0.001629001	1.629
Sums (for Doctors & Population)	440,269	336,503,000	Averages:	0.653
Doctors/Person=		0.001308366		
Doctors/Thousand Persons=		1.308	Compare as Doctors per 1,000 People	

Country	Doctors	Total Population '75		Doctors/Person	Doctors Per 1,000 People
Mexico	31,556	59,204,000		0.000533005	0.533
USA	348,484	213,925,000		0.001629001	1.629
Sums (for Doctors & Population)	380,040	273,129,000	Averages:		1.081
Doctors/Person=		0.00139143			
Doctors/Thousand Persons=		1.391		Compare as Doctors per 1,000 People	

	Country	Total Population '75 (in thousands)	Gross Nat'l Product '78 (In Millions of U.S. Dollars)		Total Population '90 (in thousands) (Statistical Abstract of the United States, 1991 Table 1359)	GNP '90
1	Afghanistan	19,280	$2,290		15,564	
2	Albania	2,482	$1,930		3,273	
3	Algeria	16,792	$25,730		25,337	$52,194,220,000
4	Andorra				52	
5	Angola	6,394	$2,810		8,449	
6	Antigua and Barbuda				64	
7	Argentina	25,384	$53,430		32,291	$76,529,670,000
8	Armenia				3,357	
9	Aruba				64	
10	Australia	13,809	$114,780		17,037	$289,629,000,000
11	Austria	7,538	$56,450		7,644	$145,694,640,000
12	Bahamas	200	$520		249	
13	Bahrain	260	$1,500		520	
14	Bangladesh	73,746	$7,280		113,930	$23,925,300,000
15	Barbados	245	$520		254	
16	Belgium	9,846	$95,450		9,909	$153,985,860,000
17	Belize				220	
18	Benin	3,074	$740		4,674	$1,682,640,000
19	Bhutan	1,173	$90		1,566	$297,540,000
20	Bolivia	5,410	$2,700		6,989	$4,403,070,000
21	Bosnia Herzegovina				4,517	
22	Botswana	691	$490		1,224	$2,496,960,000
23	Brazil	109,730	$180,020		152,505	$408,713,400,000
24	Brunei				372	
25	Bulgaria	8,793	$28,310		8,934	$20,101,500,000
26	Burkina Faso				9,078	$2,995,740,000
27	Burma	31,240	$4,480		41,277	
28	Burundi	3,765	$650		5,646	$1,185,660,000
29	Belarus				10,257	
30	Cambodia-cf. Kampuchea					
31	Cameroon	6,433	$3,950		11,092	$10,648,320,000
32	Canada	22,801	$203,980		26,538	$543,232,860,000
33	Cape Verde	292	$80		375	
34	Central African Republic	1,790	$510		2,877	$1,122,030,000

35	Chad	3,947	$650		5,017	$953,230,000
36	Chile	10,253	$15,770		13,083	$25,381,020,000
37	China / People's Republic of China / Mainland	838,803	$219,010		1,133,683	$419,462,710,000
38	Colombia	25,890	$22,990		33,076	$41,675,760,000
39	Comoros	306	$80		460	
40	Congo	1,345	$850		2,242	$2,264,420,000
41	Costa Rica	1,994	$3,390		3,033	$5,762,700,000
42	Croatia				4,686	
43	Cuba	9,481	$12,330		10,620	
44	Cyprus	673	$1,670		702	
45	Czechoslovakia	14,793	$71,640		15,683	$49,244,620,000
46	Denmark	5,026	$54,000		5,131	$113,292,480,000
47	Djibouti				337	
48	Dominica				85	
49	Dominican Republic	5,118	$4,600		7,241	$6,010,030,000
50	Ecuador	7,090	$7,400		10,507	$10,296,860,000
51	Egypt	37,543	$16,890		53,212	$31,927,200,000
52	El Salvador	4,108	$2,760		5,310	$5,894,100,000
53	Equatorial Guinea	313	$100		369	
54	Estonia				1,584	
55	Ethiopia	28,134	$3,470		51,407	$6,168,840,000
56	Fiji	577	$900		738	
57	Finland	4,652	$34,020		4,977	$129,601,080,000
58	France	52,913	$473,030		56,358	$1,098,417,420,000
59	Gabon	521	$2,130		1,068	$3,556,440,000
60	Gambia	509	$100		848	
61	Georgia				5,479	
62	Germany				79,123	$1,766,025,360,000
63	Germany East	17,127	$94,960			
64	Germany West	61,682	$631,590		63,232	
65	Ghana	9,873	$4,160		15,130	$5,900,700,000
66	Greece	8,930	$32,430		10,028	$60,067,720,000
67	Grenada	100	$60		84	
68	Guatemala	6,129	$6,130		9,038	$8,134,200,000
69	Guinea	4,416	$1,350		7,269	$3,198,360,000
70	Guinea-Bissau	525	$120		999	
71	Guyana	791	$460		753	
72	Haiti	4,552	$1,150		6,142	$2,272,540,000
73	Honduras	3,037	$1,630		4,804	$2,834,360,000
74	Hong Kong	4,225	$15,400			$0
75	Hungary	10,534	$37,150		10,569	$29,381,820,000

76	Iceland	216	$2,130		257	
77	India	613,217	$117,520		852,667	$298,433,450,000
78	Indonesia	136,044	$45,780		190,136	$108,377,520,000
79	Iran	32,923	$55,510		57,003	$141,937,470,000
80	Iraq	11,067	$22,540		18,782	
81	Ireland	3,131	$12,280		3,500	$33,425,000,000
82	Israel	3,417	$13,760		4,436	$48,441,120,000
83	Italy	55,023	$260,940		57,664	$970,485,120,000
84	Ivory Coast / Cote d'Ivoire	4,885	$7,460		12,478	$9,358,500,000
85	Jamaica	2,029	$2,540		2,469	$3,703,500,000
86	Japan	111,120	$884,500		123,567	$3,142,308,810,000
87	Jordan	2,688	$2,370		3,273	$4,058,520,000
88	Kampuchea / Cambodia	8,110			6,991	
89	Kenya	13,251	$5,180		24,342	$9,006,540,000
90	Kiribati				70	
91	Korea					$0
92	Korea North	15,852	$17,040		21,412	
93	Korea South	34,663	$48,000		42,792	
94	Kuwait	1,085	$19,410		2,124	
95	Kyrgystan				4,394	
96	Laos	3,303	$300		4,024	$804,800,000
97	Latvia				2,695	
98	Lebanon	2,869	$3,290		3,339	
99	Lesotho	1,148	$390		1,755	$930,150,000
100	Liberia	1,708	$790		2,640	
101	Libya	2,255	$19,820		4,223	
102	Liechtenstein				28	
103	Lithuania				3,726	
104	Luxembourg	342	$4,010		384	
105	Madagascar	8,020	$2,100		11,801	$2,714,230,000
106	Malawi	4,909	$1,040		9,197	$1,839,400,000
107	Malaysia	12,093	$15,270		17,556	$40,729,920,000
108	Maldives	120	$30		218	
109	Mali	5,697	$810		8,142	$2,198,340,000
110	Malta	329	$770		353	
111	Mauritania	1,283	$420		1,935	$967,500,000
112	Mauritius	899	$850		1,072	$2,412,000,000
113	Mexico	59,204	$91,910		88,010	$219,144,900,000
114	Moldova				4,393	
115	Mongolia	1,446	$1,100		2,187	
116	Morocco	17,504	$12,890		25,630	$24,348,500,000
117	Mozambique	9,223	$2,380		14,539	$1,163,120,000
118	Myanmar					
119	Namibia				1,453	
120	Nepal	12,572	$1,580		19,146	$3,254,820,000
121	Netherlands	13,599	$128,270		14,936	$258,691,520,000
122	New Zealand	3,031	$17,700		3,296	$41,793,280,000

123	Nicaragua	2,318	$2,090		3,602	
124	Niger	4,600	$1,180		7,879	$2,442,490,000
125	Nigeria	63,049	$48,100		118,819	$34,457,510,000
126	Norway	4,007	$38,790		4,253	$98,329,360,000
127	Oman	770	$2,340		1,481	
128	Pakistan	70,560	$18,250		114,649	$43,566,620,000
129	Panama	1,678	$2,280		2,425	$4,437,750,000
130	Papua New Guinea	2,716	$1,820		3,823	$3,287,780,000
131	Paraguay	2,647	$2,660		4,660	$5,172,600,000
132	Peru	15,326	$11,440		21,906	$25,410,960,000
133	Philippines	44,437	$24,410		64,404	$47,014,920,000
134	Poland	33,841	$127,560		37,777	$63,843,130,000
135	Portugal	8,762	$19,000		10,354	$50,734,600,000
136	Puerto Rico	2,902	$8,910			
137	Qatar	90	$3,310		491	
138	Romania	21,178	$36,190		23,273	$38,167,720,000
139	Russia				148,254	
140	Rwanda	4,233	$870		7,609	$2,358,790,000
141	Saint Kits and Nevis				40	
142	Saint Vincent and the Grenadines	80	$43		113	
143	San Marino				23	
144	Santa Lucia				150	
145	Sao Tome and Principe				125	
146	Saudi Arabia	8,966	$54,200		17,116	$120,667,800,000
147	Senegal	4,418	$1,930		7,714	$5,476,940,000
148	Serbia				9,883	
149	Seychelles	60	$80		68	
150	Sierra Leone	2,983	$740		4,166	$999,840,000
151	Singapore	2,248	$7,600		2,721	$30,366,360,000
152	Somalia	3,170	$340		6,654	$798,480,000
153	South Africa	24,663	$43,760		39,539	$100,033,670,000
154	Soviet Union frmr					
155	Spain	35,433	$146,940		39,269	$432,744,380,000
156	Sri Lanka	13,986	$2,870		17,198	$8,083,060,000
157	Sudan	18,268	$5,900		26,245	
158	Suriname	422	$850		397	
159	Swaziland	469	$310		837	
160	Sweden	8,291	$87,260		8,526	$201,725,160,000
161	Switzerland	6,535	$81,930		6,742	$220,328,560,000
162	Syria	7,259	$7,820		12,483	$12,483,000,000
163	Taiwan / Republic of China	16,453	$14,890		20,435	
164	Tajikistan				5,342	

165	Tanzania	15,388	$4,130		25,971	$2,856,810,000
166	Thailand	42,093	$23,390		56,002	$79,522,840,000
167	Togo	2,248	$770		3,674	$1,506,340,000
168	Trinidad and Tobago	1,009	$3,410		1,271	$4,588,310,000
169	Tunisia	5,747	$6,010		8,104	$11,669,760,000
170	Turkey	39,882	$53,890		57,285	$93,374,550,000
171	Turkmenistan				3,658	
172	Tuvalu				9	
173	UAR United Arab Emirates	220	$12,180		2,254	$44,764,440,000
174	Uganda	11,353	$3,470		18,016	$3,963,520,000
175	Ukraine				51,711	
176	United Kingdom	56,427	$319,480		57,366	$923,592,600,000
177	Upper Volta	6,032	$880			
178	Uruguay	3,108	$5,170		3,102	$7,941,120,000
179	USA	213,925	$2,135,010		250,410	$5,456,433,900,000
180	USSR	255,038	$967,820			
181	Uzbekistan				20,569	
182	Venezuela	12,213	$39,880		19,698	$50,426,880,000
183	Vanuatu				165	
184	Vietnam	43,451			66,171	
185	Vietnam North	23,800				
186	Vietnam South	19,650				
187	Western Somoa	160	$50		186	
188	Yemen				9,746	
189	Yemen (Aden)	1,660	$780			
190	Yemen (Sana)	6,668	$2,301			
191	Yugoslavia	21,322	$46,140			$0
192	Zaire	24,450	$6,480		36,613	$8,054,860,000
193	Zambia	5,004	$2,720		8,154	$3,424,680,000
194	Zimbabwe	6,272	$3,330		10,394	$6,652,160,000

		Source: World Handbook of Political and Social Indicators, Third Edition, by Charles L. Taylor and David A. Jodice, Yale University Press, New Haven and London, 1983	***Source: World Handbook of Political and Social Indicators, Third Edition, by Charles L. Taylor and David A. Jodice, Yale University Press, New Haven and London, 1983***			***Computed from Population and GNP per capita***

Exercise Life Expectancy

Average by nation

Average by person

On The Average II:
Weighted Averages and The Unit of Analysis

or:

If
A Chicken and a Half
Lays an Egg and a Half
In a Day and a Half,
Then
How Many Eggs Does a Chicken
Lay in a Day?

That title was once a common riddle among kids, designed to sound silly, confuse the mind, and otherwise pass the time. It seems to have fallen out of the common “kids culture”, which is no great loss, but its solution actually makes a point: If you can get through things like that with your mind still intact, then maybe you can get through real-world problems, deal with data in their naturally bizarre state of arrangement, and still keep your mind intact.

I actually intend to answer the question stated by the title, but I’m going to have to work up to it. Once again the moral of the story is going to be — be careful of the units and, as far as possible, build the units into your equation.

So far, for the mean, I’ve been using the simple equation

$$\bar{x} = \frac{x_1 + x_2 + \quad + x_n}{n}$$

abbreviated

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where “x-bar” is the mean, and where x_1, x_2, and all the little x_i’s up to x_n are the n values whose average is being computed.

Properly amended, building-in the unit, that becomes

$$\bar{x}\,\text{units} = \frac{x_1\,\text{units} + x_2\,\text{units} + \quad + x_n\,\text{units}}{n}$$

or

$$\bar{x}\,\text{units} = \frac{1}{n}\sum_{i=1}^{n} x_i\,\text{units}$$

And now the equation preserves the idea that whatever units are used for “x”, the thing being averaged, the average value of x is measured in the same units. If the units are people, the mean is in people. If the unit is dollars, then the mean is in dollars.

$$\bar{x}\,\text{people} = \frac{1}{n}\sum_{i=1}^{n} x_i\,\text{people}$$

$$\$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} \$x_i$$

Now I need something called the “weighted average”. The weighted introduces another set of numbers, the weights, and associates each of the x’s with a weight. Then, for the weighted average you computed the weighted sum and divide by the sum of the weights.

$$\overline{x} = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_n w_n}{w_1 + w_2 + \ldots + w_n}$$

One use of weights is to keep count: If I throw a pair of dice, and they come up 3, 4, 10, 4, and 3, in successive throws of the dice, then I can compute the average giving each toss of the dice a weight of one:

$$\overline{x} = \frac{3*1+4*1+10*1+4*1+3*1}{5}$$

which, is, of course, identical to adding up the successive values of the dice and dividing by 5. or I can use weights to keep track of the number of 3's and 4's and 10's that came up, giving 3 and 4 a weight of two as compared to a weight of one for the 10 (because 3 and 4 each came up twice):

$$\overline{x} = \frac{3*2+4*2+10*1}{2+2+1}$$

It adds a little bit to the formula for the average, but ultimately it is more compact: I can throw the dice 1,000 times, but the average for dice is always going to be 2 times a weight, plus 3 times a weight, plus 4 times a weight, and so forth, divided by the sum of the weights:

$$\overline{x} = \frac{\sum_{i=2}^{12} i\, w_i}{\sum_{i=2}^{12} w_i}$$

That's the weighted average — as arithmetic. But I want you to use the weighted average as a kind of debater's trick — I'm not trying to debate with you or trick you: The debate is the one I carry on in my head, debating with myself to see if I know what I'm doing — tricking myself with devices that get me to see a problem from unfamiliar angles.

Now back to physicians per capita. I think / hope you understand it. Now let me try to use close attention to the units, and use the weighted average to straighten the whole thing out. But one warning: Even if this looks automatic, a "method", a procedure — use it right and it will pay you back with the right answer — even if it looks that way don't believe it. I'm using this as a debater's trick, Levine v/s Levine: The formulas do not know the answer.

So now, quickly, back to doctors and the world

Unit: Country	Attribute: Number of Doctors, 1975	Attribute: Total Population, 1975		Attribute: Doctors Per Person	Attribute: Doctors Per 1,000 People
Afghanistan	656	19,280,000		0.000034	0.034
Angola	384	6,394,000		0.000060	0.060
Argentina	48,687	25,384,000		0.001918	1.918
Austria	15,702	7,538,000		0.002083	2.083
Bahamas	161	200,000		0.000805	0.805
Bahrain	177	260,000		0.000681	0.681
Bangladesh	5,088	73,746,000		0.000069	0.069
Barbados	166	245,000		0.000678	0.678
Belgium	18,510	9,846,000		0.001880	1.880
Benin	95	3,074,000		0.000031	0.031

Now, what's the question and what's the answer: If the question is a question about a typical country or nation, conceived as a unit, with its population, with a national medical system, with a national medical education system, a national public health system, hospital system and payments system. If the question is about the results for countries and for the typical county — then you want the country as the unit of analysis and you want to average the data with respect to the country. What have we got in the table? We have a list of units, the countries, and a list of attributes in each row that describe that unit. So, using a "sample" of one country, using the ".000034" (in row one) — what is the unit? It is “Doctors per person *per country*” This datum provides evidence for one country, Afghanistan. That last “per ____”, at the bottom of the stack names the unit. And the unit in the weight matches the unit of analysis. When I take the weighted average it gets the weight “1 country”. In the second row I have attributes of another unit, another country, Angola. The .000060 in the second row is “Doctors per person per country” (for this second country). Now, using the unit of analysis as the unit of weight,

$$\overline{x} = \frac{\sum_{i=2}^{12} i \; w_i}{\sum_{i=2}^{12} w_i}$$

$$\overline{x} = \frac{\left[0.000034 \frac{\frac{doctors}{person}}{country}\right] * 1country + \left[0.000060 \frac{\frac{doctors}{person}}{country}\right] * 1country + \left[0.001918 \frac{\frac{doctors}{person}}{country}\right] * 1country + \ldots}{138 countries}$$

I get the average of _____ $\left(\frac{\frac{doctors}{person}}{country}\right)$ And what does this thing,. this average, describe: Look at the units on the average, and there it is, at the bottom of the stack: the country, the average country.

Now, by contrast, suppose I had done it the other way — adding up the numbers in the column for doctors: At the bottom of the data sheet I show a sum of 2,622,088 *doctors*, two-point-six million doctors. And, this way, what is the unit? What do these doctors belong to? What unit has two-point-six million doctors? The unit is the world itself (or the reporting nations). In the next column the table shows 3,028,196,000 people. And what do these people belong to? The same world. And their ratio describes the world: In the world there is a ratio of .000865891 doctors *per person*.

And now, I have to deal with a "chicken and a half" So, what is my unit? What does an egg and a half in a day and a half describe? An egg and a half in a day in a half is an attribute describing 1.5 chickens. — my unit is the chicken, (although I'm given data about a chicken and a half). So the attribute that describes the chicken and a half is

$$\frac{1.5\ eggs}{1.5\ days}$$

And, putting the rate in the proper relation to the unit it describes, as the rate for a chicken and a half: [1]here is the "formula" in eggs per day per chicken:

$$\frac{1.5\ eggs}{1.5\ days}\Big/1.5\ chickens$$

The units and the unit of analysis are clear: eggs per day *per chicken*. So I just clean up the pesky fractions and get one egg per day for 1.5 chickens:

$$1\frac{egg}{day}\Big/1.5\ chickens$$

or two-thirds of an egg per day per chicken

$$\frac{2}{3}\left(\frac{eggs}{day}\Big/chicken\right)$$

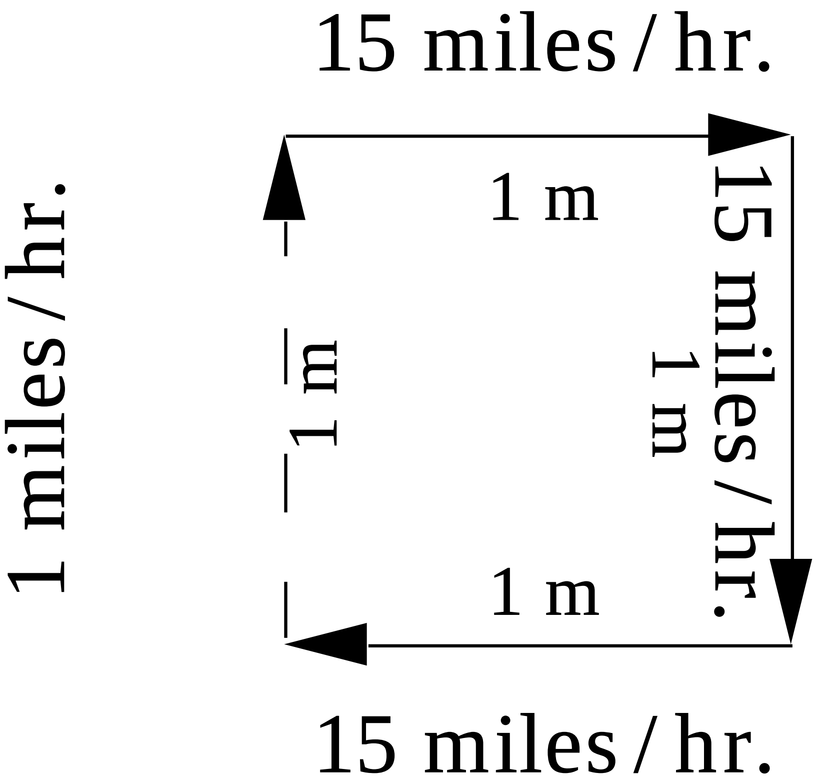
That is: One chicken, the unit, lays two-thirds of an egg per day. Watch the unit of analysis, watch the chicken, not the egg (or the day).

[1] Exercise for the reader: How many days does it take for a chicken to lay an egg? (Solve for x, whose unit will turn out to be a number of days, in the equation $\frac{2}{3}\left(\frac{eggs}{day}\Big/chicken\right) * x = 1\ \frac{egg}{chicken}$

On the Average III: Racing

O.K., you have the principle: Watch the unit of analysis. You have the principle under control, but let me put your control to another test with Data Gymnastics exercise #2:

Consider a foot race on a 4 mile square track, 1 mile on an edge. I'm racing against myself, going for a personal record — and then some.



Here's my performance: I managed 15 miles per hour for the first mile (a four minute mile), I repeated at 15 miles per hour for the second mile, repeated again at 15 miles per hour for the third mile. And then I paid for the first three miles with heat stroke, plus a heart attack and sore feet — leaving me able to crawl the last mile in exactly one hour. What's was my average?

Well, it was

$$(15 + 15 + 15 + 1) / 4 = 11.5$$

If that looks reasonable to you, if 11.5 miles per hour seems like a reasonable estimate for the average, then explain this: I ran and crawled for a little more than an hour, for exactly one hour and twelve minutes. So how is it that I covered a distance of only four miles, which is considerably less than the 11.5 miles that I am supposed to have traveled in an average hour? That makes no sense. So, the answer, 11.5 miles per hour, must be wrong.

The way out of this anomaly is, again, to watch the unit of analysis. Does your average, 11.5, describe the typical mile: What I accomplished during the first mile averaged with what I accomplished in the second mile, and the third, and the fourth. Or does your average describe the typical hour or minute: Averaging what I did in the first four minutes with what I did in the next four minutes — with what I did in the next four minutes, and what I did in the next sixty minutes? What is the unit of analysis? Is it a mile, or is it a minute? It matters.

That is the nature of the puzzle for this bit of mental gymnastics, but let me tell you how I really figure out the solution to a problem like this: The rule for problem solving in this situation is another one of those strategies that you never admit to in a final report. The rule is work backward. Create a simple thought experiment: Figure out the answer for a hypothetical example, *then* figure out the method that works for the hypothetical example, and then go back to the data where you can solve the problem. What you do not want to do is just forge ahead, however bravely, with “2 unknowns”: Trying to use an unknown method in search of an unknown answer. So you do a thought experiment. You think up a problem for which the answer is known. You work on this problem until the method becomes clear, and then you go back to the data.

So, working backward, what do I know about this problem? Actually, working backward, I can begin with the answer: I know that the total mileage was four miles and that the total time was 1 hour and 12 minutes, or 1.2 hours. So the answer is going to be 4 miles per 1.2 hours, which is 3.33 miles per hour. I have the answer. Now, what’s the method? What type of average would have given me the right result?

Well, the method is not what I did above, not (15 + 15 + 15 + 1)/4 = 11.5, but let’s take a look at it using the units and using the weighted mean. What went wrong?

$$\text{mean} = \frac{15\frac{\text{miles}}{\text{hour}} * 1\text{ mile} \quad + \quad 15\frac{\text{miles}}{\text{hour}} * 1\text{ mile} \quad + \quad 15\frac{\text{miles}}{\text{hour}} * 1\text{ mile} \quad + \quad 1\frac{\text{mile}}{\text{hour}} * 1\text{ m}}{4\text{ miles}}$$

If you look at that in detail, it makes no sense: the units used as the units of analysis are inconsistent (and that's what got me the wrong answer): When I wrote 15 miles *per hour*, I have a unit of time where I should expect the unit of analysis. But when I wrote 1 *mile* for the weight I said that I was organizing the analysis in miles. I can't do both (not at the same time). I have to pick one. In fact, I can organize the problem in terms of time, or I can organize it in terms of distance — but which ever I do I have to do it consistently.

Lets try it in terms of hours: I started out at 15 miles/hour. Hours are in the denominator so I'll give this speed a weight in hours. Thinking it through, I kept up this speed for four minutes in order to complete the first file. Four minutes is one-fifteenth of an hour — there's the weight. I did it again, same speed, for the next one-fifteenth of an hour, completing the second mile. I did it again, same speed, for the next one-fifteenth of an hour, completing the third mile. And then I completed my circuit at one mile per hour, continuing at that rate for an entire hour until I was finished. Using the hour as the unit, here is a consistent weighted average:

$$\text{mean} = \frac{15\frac{\text{miles}}{\text{hour}} * \frac{1}{15}\text{ hour} \quad + \quad 15\frac{\text{miles}}{\text{hour}} * \frac{1}{15}\text{ hour} \quad + \quad 15\frac{\text{miles}}{\text{hour}} * \frac{1}{15}\text{ hour} \quad +}{1\frac{3}{15}\text{ hours}}$$

which simplifies to

$$\text{mean} = \frac{1\text{ mile} \quad + \quad 1\text{ mile} \quad + \quad 1\text{ mile} \quad + \quad 1\text{ mile}}{1.2\text{ hours}}$$

and to

$$\text{mean} = \frac{4 \text{ miles}}{1.2 \text{ hours}}$$

which gives

$$\text{mean} = 3.33 \, \frac{\text{miles}}{\text{hour}}$$

which is correct.

There's my reward for thinking clearly. Now I've got the formula for the right answer. I've thought it through. I've been very careful to keep the unit of analysis simple and in the right place, and I got the right answer. The moral: Watch the unit of analysis. Check, with a simple problem — just to be sure. And understand that averages, in the real world, require patience and careful thinking, — even for "easy" problems.

Exercise: I picked one unit, the hour. But suppose I had picked the mile as the unit, re-expressing my first mile as 4 minutes *per mile* (and my fourth mile as 60 minutes*per mile*)? Now I am organizing the analysis in miles and I need a weight in miles. Will it work? (Partial answer: It can't give me exactly the same answer because the first answer was in miles per hour and the second answer will be in hours per mile (or minutes per mile). Keeping that in mind, keeping in mind that the correct answer in hours per mile is exactly the inverse of the correct answer in miles per hour — the answer will be *almost* the same and it will be correct. Do it and explain.)

Homework. Most of the hard work of data analysis is undertaken flat on your back, staring at the ceiling, with your eyes closed — at least for me. Whatever your equivalent posture, assume it: You won't need much computing for this one, except to write down your answers:

#1: I've seen versions of this in magazines for people who get a little nuts about precise measurements of aerobic performance and measures of improvement (or change) from day to day:

Question: Dear Dr. C.: I understand the rules for measuring my aerobic capacity. But living in hill country, far from a track, how do I measure my performance when my route takes me up hills and down?

Answer: Run a circular route. For every mile you have to run uphill you will be treated to a mile downhill and so the effort and time, on a flat route, are equivalent to average effort and average time on a circular route — no adjustments necessary.

Problem: Obviously Dr. C. lives in flat country, otherwise his heart would long ago have punished his brain for a nice simple logical idea — that is wrong. Explain to the poor man why his averaging doesn't work, lest his mind kill his body on their first trip to the mountains. Create a simple numerical example, using a weighted average to illustrate the problem.

On the Average IV:
Money — How am I Doing?

This one puzzles me too. But it's real.:

The immodest truth is that data analysts can solve any problem involving numbers. And, to prove the point, a few months ago I decided to solve one of the outstanding problems of the world — getting rich. I analyzed the situation thoroughly, spending several hours of my time on it. And I figured it out. However, I realized that some people might be skeptical of my solution, were I to settle for the intellectual satisfaction of being right: After all, on paper it's one thing, but show me your track record. So to prove my point, I invested $4,000 three months ago. Now, three months later, I've got a track record. So send all your money to my mutual fund and I'll take care of it for you.

The track record? Oh, here it is: My strategy is to buy stocks in an industry that has a small probability of a very large gain: I won't win often, but when I get the right industry I'll win big — or at least enough to offset the losses suffered when I guess wrong. My theory of the economic cycle narrowed the selection down to four industries, but I didn't know which one of the four would be the winner. I knew it would be high tech, banking, gold, or transport, but I didn't know which one. So I sampled from each of the four industries: I bought shares of 10 stocks in each industry, forty stocks in all. And I waited to see what happened.

What happened was — it worked: I invested $1,000 in high tech, risking $100 in each of ten companies. And the high tech stocks went up:

The average high tech company gained 25% in the first quarter. That gave me

$1,000	+	25% of $1,000	=	$1,250	after the first quarter	(+25%)

So, extrapolating forward, I project that by the end of the year I will have $1,464.10 of high tech stocks. That's,

$1,000	+	25% of $1,000	=	$1,250.00	after the first quarter	(+25%)
$1,250	+	25% of $1,250	=	$1,562.50	after the second quarter	(+56%)
$1,562.50	+	25% of $1,563	=	$1,953.13	third quarter	(+95%)
$1,953.13	+	25% of $1,953	=	$2,441.41	fourth quarter	(+144%)

My three other industries are lagging but, as I said, I was after one big winner that would carry the rest. I didn't know which industry it would be — it turns out to have been high tech. By contrast, the $1,000 in banking lost 10 percent in each of the first two quarters. That brought the $1,000 invested in banking down to $900 by the end of the first quarter and down to $810 by the end of the second

$1,000	–	10% of $1,000	=	$900	after the first quarter	(-10%)

And by the end of the year it will be

$1,000	–	10% of $1,000	=	$900	after the first quarter	(-10%)
$900	–	10% of $900	=	$810	after the second quarter	(-19%)
$810	–	10% of $810	=	$729	third quarter	(-27%)
$729	–	10% of $729	=	$656.10	fourth quarter	(-34%)

The same relatively small losses happened to the gold and the transports. So by the end of the year my $4,000 investment will be worth $4,409.40, which is a handsome performance, up 10.2%

That's

	$2,441.10	High Tech
plus	656.10	Banking
plus	656.10	Gold
plus	656.10	Transportation
equals	$4,409.40	Total

Summarizing these computations in tabular form:

Industry:	Initial	First Quarter	Second Quarter (projected)	Third Quarter (projected)	Fourth Quarter (projected)	*Gain/Loss*
High Tech	$1,000	$1,250	$1,563	$1,953	2441	*$1441*
Banking	$1,000	$900	$810	$729	$656	*-$344*
Gold	$1,000	$900	$810	$729	$656	*-$344*
Transport	$1,000	$900	$810	$729	$656	*-$344*
					Average Gain:	*$102 (10%)*

So, send me your money: Ten percent return, almost guaranteed.

Further Correspondence

I have received a letter of complaint — somebody who doesn't understand what I've done. He works at something called "SEC", whatever that is. Here's what he said:

Dear Jail bait:

You don't know how to compute an average. Here is your real track record: During the first quarter you didn't gain money, you lost. And your average stock lost $12.50. Here are the numbers

High Tech:	Plus	$250
Banking:	Minus	$100
Gold:	Minus	$100
Transportation:	Minus	$100
Average Gain:	Minus	$12.50 (*negative* 1.25%)

Your average stock lost $12.50 in the first quarter. That's not good. With $1,000 invested in each stock, you lost an average of 1.25% So, roughly speaking (ignoring the compounding of the interest), your portfolio is down 1.25% for the first quarter. It will be down about 2.5% for your first six months. And you can anticipate loosing 5% for the year.

That was the letter. Clearly this person understands a lot less than meets the eye. Good thing he works at SEC instead of handling money. So as an educator, always eager to spread truth and reason, I respond.

Reply

Dear Sound and Fury:

Thank you for your letter. You are dead wrong, but I appreciate your comment and the opportunity to explain. And, more practically, you have pointed out how I can make substantially more money than I had proposed.

The problem with your dismal forecast is that you failed to use the right unit of analysis. You used the company as the unit of analysis and averaged with respect to all stocks. I used the industry as the unit of analysis because my theory predicts industries and that’s the theory I’m testing. I analyzed my data by industry, using four separate samples of corporations to compute my four separate averages for four separate industries. Remember: I knew from my theory that one or more of the four industries, either high tech, banking, gold, or automobiles, would be a winner. But I didn’t know which one. So I selected a sample of companies in each industry and waited. And after three months my best estimate of each industry’s performance is based on the average performance of my high tech stocks, representing the industry. Then, six months later, data in hand, my estimates for high tech was “Up”, while my estimates for each of the other industries was “Down”.

And the good part, for which I thank you, is that now that I know which industry is going up I can sell my shares in the other three — the other shares have served their purpose. I can put all that money into high tech industry, which is going up. If present trends continue, I'll make a bundle. I hope you will join me.

Your error, if I need to explain further, was to compute the average across the portfolio. That's wrong. That's like averaging apples and oranges (or, in this case, like averaging high tech and banks) and is just plain wrong. By contrast, using the procedure specified by my theory, I have four separate problems and four separate averages: And now, by industry, I know that the average high tech company is going up. Based on my four carefully separated analyses of these four separate industries: High tech is going up, banking, gold, and transport are going down. I urge you to join me, while the moment lasts.

To the Student:

Exercise: "Discuss." As far as I can tell either explanation could be correct — or made correct with a little more polish. I doubt that the Securities and Exchange Commission would pursue either line of reasoning as fraudulent. And that's distressing because these two opposite evaluations are based on one set of facts. Don't send your money, but be sure you understand the arguments.

On The Average V:
Class Size

For my fourth data gymnastic, it's time to be practical. This is one that touches the experience of every college student: You are a college student, or a prospective college student. And one of the things that's important to you in evaluating a college is the average class size. I suppose you could always ask the college about its class size, or look it up in a published ranking of colleges, or read the number in the catalogue. But let's suppose you had to compute the number yourself — just to leave nothing to chance and to eliminate confusion. That's the way to understand it — do it yourself, at least on a hypothetical example.

O.K., let's start with another thought experiment — something too simple for reality, but clear enough for practice. Suppose my college has exactly two hundred students and one class — everybody takes it. With once class and no variation, that's easy, at least for the average: There is exactly one class, everyone has the same class, and the average class size is 200.

That's the simplest example. Now for something one step up in difficulty: My college just decided that some of the seniors need special treatment, perhaps some lab. work. Now there are two classes, one of size 190, one of size 10. What's the average class size? Obviously, the arithmetic is (190 + 10) / 2 = 100. The average class size is 100.

Class 1	190
Class 2	10
Sum	200
Average	100

Let me push this a little further, trying to create something more interesting: It also turns out that we have two students with really special needs, and we have a faculty willing to accommodate. So, now there are four classes, one of size 190, one of size 8, and two of size 1. What's the average? Obviously, the arithmetic is (190 + 8 + 1 + 1) / 4 = 50.

Class 1	190
Class 2	8
Class 3	1
Class 4	1
Sum	200
Average	50

Now, I ask you a question: Do you "believe" that answer, "50"? I'm not asking whether or not my arithmetic is correct. It is. I'm asking whether or not you would feel you had been cheated if this college had advertised: "Our average class size is 50." and you had chosen to attend? If you believe this answer, then let me give you an example that is more extreme: This time I'm going to give those ten seniors individual treatment, one on one with their professors. Again, what's the average? The arithmetic would seem to be (190 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1) / 11 = 18.18, or 18 (approximately).

Class 1	190
Class 2	1
Class 3	1

Class 4	1
Class 5	1
Class 6	1
Class 7	1
Class 8	1
Class 9	1
Class 10	1
Class 11	1
Sum	200
Average	18.18

Still satisfied? If you find that number acceptable, then consider: Suppose you went to this college, with an advertised average class size of 18. And suppose that, somehow, now that you're there, the experience isn't exactly living up to your expectations. So one day at lunch you decide to check for yourself and you starting asking the other students: "My class has 190 students in it. How large is yours?" And sure enough, nearly all of the other students at the lunch table report that their class too has 190 students in it (in fact, you're all in the same class).

Suitably alarmed, but recognizing that your lunch table crowd may not represent the whole college, you decide to check further: You decide that it's time for a survey and so you send a questionnaire to every student: You want the facts. What's the size of your class?

And here's how the numbers come back:

Ten students report classes of size one. and one hundred and ninety students report classes of size 190. So, using your own data, what's the average? Well, you got two hundred responses to your questionnaire and so the arithmetic requires you to add up the numbers and divide by 200. That's

Student #1	1	
Student #2	1	
Student #3	1	10 students
...		
Student #10	1	
Student #11	190	
Student #12	190	
Student #13	190	190 students
...		
Student #200	190	
Sum	36,110	200 students
Average	180.5	

According to your survey, the average class size is 181 ! The college says the average class size is 18. But you asked the students and your data say that the average class size is 181 — ten times larger than advertised.

The Unit of Analysis (Again)

What's the problem? Don't say the problem is that, "Statistics lie," or that "The average (one of them) isn't the right number." Neither of these epithets is an intelligent response to a confusing situation — These are just different ways of re-stating the fact that there's a problem. But what *is* the problem? How can there be two radically different answers, both 18 and 181, as answers to the question "What's the average?"

Actually, both answers are correct. And both the college, that reports 18, and you, with your survey that reports 181, are using the same facts.

In this case, again, you straighten out the mess by asking yourself what is the *unit* you are interested in? What is the *unit of analysis*? If I were a prospective college student, looking out for my own best interest (defined as small classes), then I would take the student as the unit of analysis. I would look at the data that describe each unit, as you did in your survey, and then I would average them. And, just to be sure I got the thing right I would identify the units, as well as the numbers, when I wrote down the equation.

So, unit number one, that is *student* number one, reported 190 *classmates* in class. So, with respect to the student as the unit of analysis (look at the denominator), the first report is

$$190\ \frac{\text{classmates}}{\text{student}}$$

Read that as "190 classmates *per* student". And, to repeat, note the denominator and note the use of labels, "classmates" and "students"

Student number two came up with the same report, and now I have two instances of 190 classmates per student. And that would repeat for 190 of the 200 reports.

$$\left.\begin{array}{l} 190\,\dfrac{\text{classmates}}{\text{student}} \\ 190\,\dfrac{\text{classmates}}{\text{student}} \\ \text{....} \\ 190\,\dfrac{\text{classmates}}{\text{student}} \end{array}\right\} \quad \text{190 reports}$$

And then for ten more students the report would be one classmate per student.

$$\left.\begin{array}{l} 1\,\dfrac{\text{classmate}}{\text{student}} \\ 1\,\dfrac{\text{classmate}}{\text{student}} \\ \text{....} \\ 1\,\dfrac{\text{classmate}}{\text{student}} \end{array}\right\} \quad \text{10 reports}$$

So, altogether, with 200 reports I have 190 reports of “190 classmates per student” and 10 reports of “1 classmate per student.” So adding up the 200 reports and dividing by 200: the average number of classmates, averaged over the 200 units, is 180.55 classmates *per student*.

And that *is* the correct answer — *for the student as the unit of analysis*. So where did the college’s advertisement of 18 students per class come from? The answer is in the denominator: students *per class*? The college used a different unit: Trust me as a faculty member — students are not the only constituents of a college: There is an administration that sweeps the floors and there are faculty that teach the classes. And, like you, both of them ask: “How big are the classes? But they get a different answer. And there’s the problem: We’re not all asking the same question of these data. The English is sloppy enough to make it look the same. But you have to look very closely at the question: As a faculty member, or taking the faculty as the unit of analysis, there are exactly eleven units in these data, eleven classes.

Attaching the units to the numbers, and spelling it out in detail, I have one faculty member who reports 190 students. That is

$$190\,\frac{\text{students}}{\text{faculty}}$$

And I have ten happy faculty reporting one student per faculty member. So my data are

$$\begin{array}{ll} 190\,\frac{\text{students}}{\text{faculty}} & \text{1 report} \\ \left.\begin{array}{l} 1\,\frac{\text{student}}{\text{faculty}} \\ 1\,\frac{\text{student}}{\text{faculty}} \\ .... \\ 1 \end{array}\right\} & \text{10 reports} \end{array}$$

And now, applying my arithmetic to these eleven numbers, one datum for each unit of analysis, I add up the numbers and divide by eleven, getting the answer eighteen, approximately, eighteen, not one hundred and eighty one.

And the moral is: Keep your eye on the unit of analysis that shows up, in this example, as the unit of the denominator — as something *per something*, where that last *something* is the unit of analysis. There are two answers here, referring to two different questions, and to two different units of analysis. Both answers are *numerically* correct, but that is not to say that both of the *answers* are correct: It all depends on the question you asked (or intended to ask — now that you are being more careful).

Writing it as a weighted mean, as

$$\text{mean} = \frac{\text{number} * \text{weight} + \text{number} * \text{weight} + \ldots + \text{number} * \text{weight}}{\text{sum of weights}}$$

in this example, using the class as the unit of analysis, each report had an implicit weight of 1, each class was given equal weight. But I can simplify the arithmetic by using the weight to represent the frequency of each value that was reported — using a weight that represents a number of the units of analysis. Thus, more efficiently

$$\text{mean} = \frac{190\,\frac{\text{students}}{\text{class}} * 1\text{class} + 1\,\frac{\text{student}}{\text{class}} * 10\text{ classes}}{11\text{classes}}$$

giving me the 18.18 students per class.

Exercise: For practice, using weight for the average of classmates per student, ...

Exercise: Consider the same problem, using medians

Exercise: It would be reasonable to assume that either one of these numbers will do: That which ever average you use, you will still get the same rank order of "best" colleges, that the best, measured in terms of classmates per student will still be the best in terms of students per class — even though the numbers will be different. Question: Is this assumption correct? When you compare two colleges, is it necessarily true that the college with the smallest number of students per class is also the college with the smallest number of classmates per student? Prove that this is not true by creating a counter example. Construct hypothetical data for two colleges such that one of the two has the smaller number of students per class while the other has the smaller number of classmates per student.

Homework:

I am enclosing three Excel files: These are enrollments during one quarter for all classes at Dartmouth. For your joy/curiosity/whatever I have included the entire data set, by class, as one file. These were the enrollments for Winter of 1991. For your homework, I have extracted the data for two departments. Practicing on something small — make up something even simpler until you're sure of your self. A hypothetical department with only two classes. Then, and only then, work yourself up to data: Take a look at Physics and take a look at Government. In each case compute an average with the class as the unit of analysis — computing the average number of students per class — and compute an average with the student as the unit of analysis — computing the average number of classmates per student. Which department, Physics of Government, has the "best" average. If you've got the numbers right, then it will not be easy to answer that question. Explain, briefly, in English — pass the room mate test for clarity and succinctness. (And what did the college tell you about class size?) (Feel free to compute the averages for the college as a whole. You have the data, at least for one term.) (Hint: You can check your computations, or mine: For Physics I get an average of 18.84 students per class and an average of 86.29 classmates per student. Get that right, or correct me, and you are ready to tackle Government — and the whole.) Do you think that the rank order of departments by class size would change, depending upon the measure used to report "size"? Could the rank order of colleges change, depending upon the measure used to report "size"?

Class							
Department	Number	Section	Type	Time	Prof.	Limit	Enrollment
GOVT	5	1	LEC	11	Masters, Roger D	105	91
GOVT	6	1	LEC	9S	Arseneau, Robert B	105	89
GOVT	7	1	LEC	11	Becker, David G	40	45
GOVT	32	1	LEC	11	Winters, Richard F	50	42
GOVT	42	1	LEC	12	Kopstein, Jeffrey S	50	63
GOVT	48	1	LEC	2A	Lustick, Ian S	999	21
GOVT	57	1	LEC	10	Becker, David G	999	69
GOVT	62	1	LEC	11	Mather, Lynn M	60	60
GOVT	64	1	LEC	9L	Masters, Roger D	999	32
GOVT	70	1	LEC	10	Sullivan, Denis G	999	23
GOVT	80	1	LEC	ARR	Winters, Richard F	999	16
GOVT	84	1	LEC	3A	Mather, Lynn M	999	17
GOVT	86	1	LEC	2A	Sa'adah, M Anne	999	11
GOVT	99	1	LEC	3A	Becker, David G	999	10

(((Class)))							
Department	Number	Section	Type	Time	Prof.	Limit	Enrollment
PHYS	4	1	LEC	11	Thorstensen, John R	999	128
PHYS	13	1	LEC	10	Mook II, Delo E	999	113
PHYS	16	1	LEC	9L	Montgomery, David C	999	30
PHYS	24	1	LEC	11	Walsh, John E	999	12
PHYS	44	1	LEC	9L	La Belle, James W	999	14
PHYS	66	1	LEC	10	Denton, Richard E	999	10
PHYS	72	1	LEC	ARR	Lawrence III, Walter E	999	4
PHYS	82	1	LEC	ARR	Sturge, Michael	999	6
PHYS	85	1	LEC	ARR	Harris, Joseph D	999	2
PHYS	87	1	LEC	ARR	Harris, Joseph D	999	2
PHYS	102	1	LEC	10A	Yafet, Yako	999	6
PHYS	105	1	LEC	11	Hudson, Mary K	999	6
PHYS	110	1	LEC	ARR	Lotko, William	999	2
PHYS	121	1	LEC	ARR	Harris, Joseph D	999	0
PHYS	122	1	LEC	ARR	Yafet, Yako	999	0
PHYS	123	1	LEC	ARR	Harris, Joseph D	999	0
PHYS	127	1	LEC	ARR	Harris, Joseph D	999	0
PHYS	137	1	LEC	ARR	Harris, Joseph D	999	11
PHYS	157	1	LEC	ARR	Harris, Joseph D	999	12

164 Enroll91WAll is in the Excel folder:

Enrollment for all classes

Volume II: Lines

analyze … to separate or break up (any whole) into its parts so as to find out their nature, proportion, function, relationship, etc.

— *Webster's New World Dictionary of the American Language, College Edition*

Lines

The prima donna of models in the social sciences is the simple straight line. Everyone knows what a line is: In a sketch it is just a trace on a graph — it has a direction and a certain height on the page. In geometry you learn that a straight line is determined by two points. In algebra you learn to match the geometry with an equation for the line, y = mx + b. In data analysis it says that the value of a variable, "y", is proportional to the value of a variable "x", with a constant "b" added to the result. And here it is: a straight line, drawn on graph paper and described with algebra.

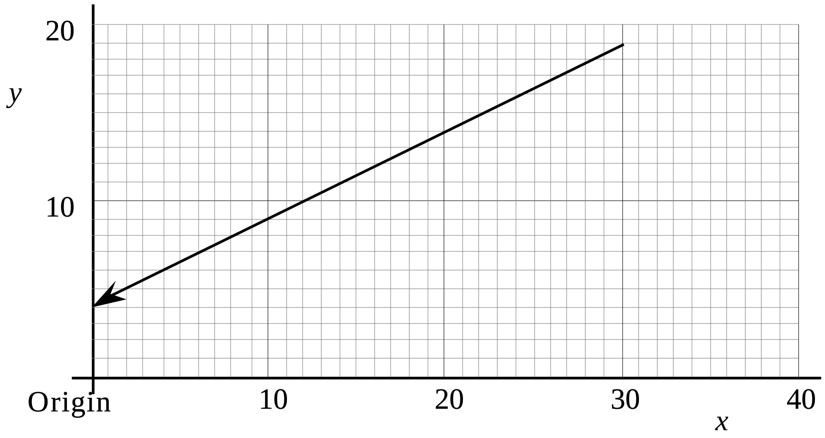


Figure __
A "mathematical" line, y = mx+b, b =4, m = .5

The slope of the line is “m”, meaning that if the interval between two values of “x” is one, drawn horizontally, then the interval between the two corresponding values of “y” is m, drawn vertically. The intercept is “b”, meaning that if the value of x is 0 then the corresponding value of y is b. b is the height at which the line *intercepts* the vertical axis.

Using data, outside of mathematics, the line is our most-used way of fleshing out the detail by which one thing is related to another. If I describe the economic consequences of education by saying, “Comparing adult Americans, every year of formal education corresponds to an addition of three thousand dollars to the average income,” that description is a line with a slope of three thousand dollars per year. The intercept is unspecified by this description, but if it were zero it would mean that no education implied no money.

In much of the sciences, for descriptive work where theory is lacking, we have nothing comparable to the decidedly non-linear ellipses of planetary motion, the quadratics of acceleration, or oscillations of springs found in elementary theoretical physics. But that’s it — we use lines. On the other hand, using logs, and any other transformation that can help, we interpret “lines” so broadly that even the quadratic equations of velocity and acceleration can be handled *indirectly* by “linear” technique.

In data analysis it is rarely the case that we even have a continuous trace on a graph, straight or otherwise. In fact it is useful to amend one of the rules that mathematicians use for lines: In math, two things, the intercept and the slope, tell you everything there is to know about a line. By contrast, in data analysis, you need the intercept, the slope, and the scatter. Usually the best we have is a series of observations that line up (more or less)

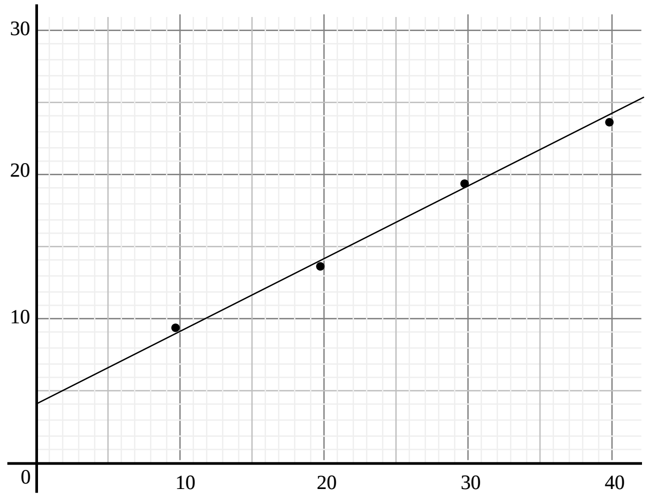


Figure A

And often what we have on the graph is less like a mathematically fine trace and more like an ambiguous cloud within which we may try to detect a pattern whose general shape may, possibly, and to some degree, be described by a line.

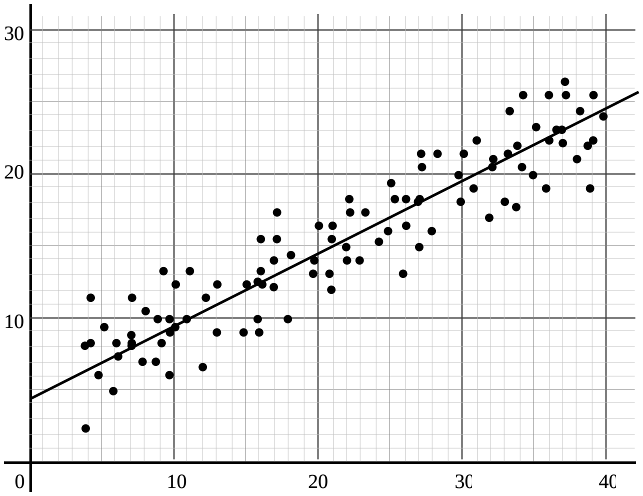


Figure B

It gets worse, much worse — the intercept, the slope and the "scatter", usually a lot of scatter — leading to such aphorisms as "You only need strong statistics when you've got weak data." If you are going to draw a graph describing 10,000 people, recording their numbers of years of education and their dollar incomes, then the folk wisdom — that there are people who never went to school, and got rich, while there are others who spent their lives in school, and made no money — will become a reality on your graph. You will have "points" (representing people) spread across all possible combinations of education and income. And it does no good to protest that the millionaire dropout and the impoverished Ph.D. don't count. It does no good to protest that these are errors, or exceptions, or deviations. Such protests are like the complaints of biology students doing their first dissection of a frog: "Things aren't where they are 'supposed to be'". No, things *are*. There is no *supposed to be* — there is nothing "defective" about the real world. Abstractions and averages are extremely valuable, indispensable really. But, no, sorry, the data are the reality. Reality is sovereign, not ideas. — This is what makes the line of the data analyst different from the line of the mathematician.

Beginning at the Beginning

Data analysis can get very complicated, at least as complicated as the world we try to discover through the analysis. That means you must proceed with caution. Simply launching into the data, drawing graphs, estimating lines, looking for correlations, letting your computer show off all the options of which your software is capable is no way to begin. You will simply overwhelm yourself with the possibilities, generate a mess, and quite possible deceive yourself — either by thinking you have found things that are not there or, more likely, by failing to find things that are.

So for "multivariate analysis", begin at the beginning or, as suggested for one variable analysis, begin before the beginning with "Who, What, Where, Why, When, and How?".

Then the beginning for two variable is one variable analysis. These variables are the building blocks for two variable analysis with and they've got to be right — if they're not right you needn't bother with the rest. So, patience: Stem and leaf, looking for tell-tale patterns, looking for outliers and above all, looking for well-behaved variables.

Then, and only then, two variable analysis. And two variable analysis begins with a well-labeled graph. Very few computer software packages will give you a well-labeled graph. By well-labeled graph I mean something more than neat and pretty. By a well-labeled graph for two variables I am suggesting something comparable to the stem and leaf diagram for a single variable. It should organize the data so that my eyeball can look and so that my intuition can see. And what I am looking for is patterns. The graphs are so important that if I had to bet on results from two different analysts, one with a computer and conventional but inadequate software, the other tackling the data with graph paper, a pen, a ruler, and hand computation — I would bet on the second analyst.

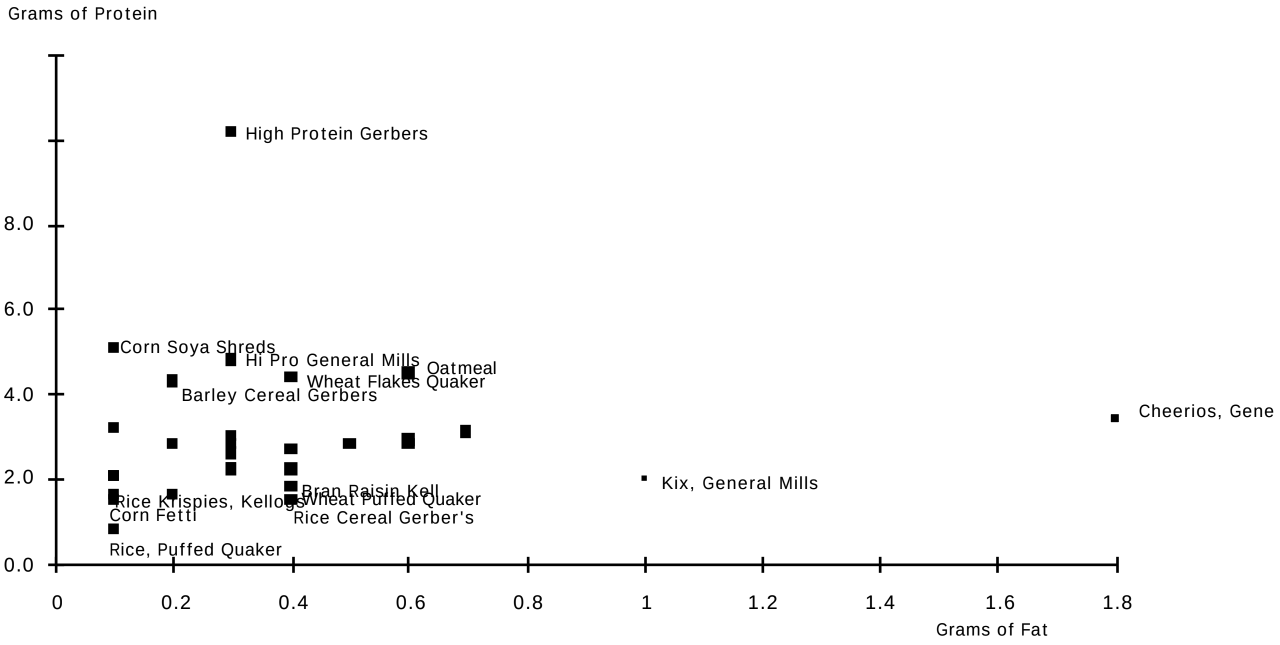
Graphs

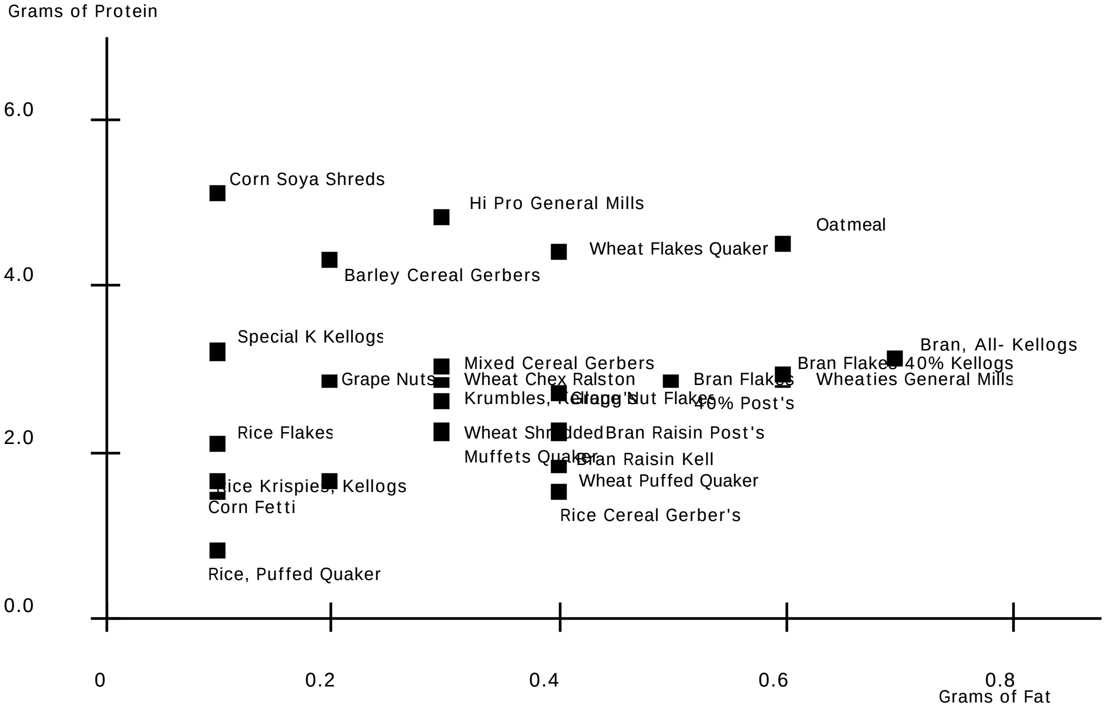
Figure 1 shows a two variable relation extracted from the breakfast cereal data. Beginning with each of the two variables, for a one variable analysis, each is close enough to being symmetrical. For Fat the median is 0.3 grams, the mid quartile is lower at 0.25 grams, and the mid-eighth is higher at 0.35 grams. For Protein the median is 2.75 grams, the mid quartile is lower at 2.65 grams, and the mid eighth is higher at 3.025 grams.

So, having completed the preliminaries I can proceed to the graph, What I see is outliers that are immediately apparent. I recognize the high protein cereal from previous work. I see Cheerios and Kix which sets me thinking about the bran component of cereals. I will forego the pleasures of another detailed analysis of breakfast cereals, except for two points.

The first point is the labels: The labels on the points are the ones I think of when I refer to a well-labeled graph. As they did in a stem and leaf, these labels feed my intuition by allowing me to connect the data to other knowledge, other knowledge carried by the label itself, other knowledge that I have in my own experience and can connect to those labels. It is, I'll admit, a "pain" to draw such labels and even here I had to draw a second graph, a close-up, to get it done. It is difficult here with only 30 points. It is very difficult with more data points. But I have no need to be mechanical or fair to these data — I'm preparing the data to feed my intuition — so I will label outliers, I will label a few familiar points, I will label points that correspond to my hunches — pursuing the possibilities as is my purpose at the beginning of the analysis.

The second point is the relation itself: This pattern, overall, is hardly what I would call a line and that is an interesting result. It may be a weak correlation but that is very useful information. It is useful to know that the two nutrients occur relatively independent of one another.


Grams of Protein
High Protein Gerbers
8.0
6.0
Corn Soya Shreds
Hi Pro General Mills
Oatmeal
Wheat Flakes Quaker
Barley Cereal Gerbers
4.0
Cheerios, Gene
2.0
Kix, General Mills
Bran Raisin Kell
Rice Krispies, Kellogs
Wheat Puffed Quaker
Corn Fetti
Rice Cereal Gerber's
Rice, Puffed Quaker
0.0
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
Grams of Fat


Grams of Protein
6.0
Corn Soya Shreds
Hi Pro General Mills
Oatmeal
Wheat Flakes Quaker
Barley Cereal Gerbers
4.0
Special K Kellogs
Bran, All- Kellogs
Mixed Cereal Gerbers
Bran Flakes 40% Kellogs
Grape Nuts
Wheat Chex Ralston
Bran Flakes
Wheaties General Mills
Krumbles, Kellog's
Grape Nut Flakes
40% Post's
2.0
Rice Flakes
Wheat Shredded
Bran Raisin Post's
Muffets Quaker
Bran Raisin Kell
Wheat Puffed Quaker
Rice Krispies, Kellogs
Corn Fetti
Rice Cereal Gerber's
Rice, Puffed Quaker
0.0
0
0.2
0.4
0.6
0.8
Grams of Fat

Foregoing a deep and detailed analysis of breakfast cereals, in this case the initial two variable picture tells me little that is not present in the one variable information. The two variable picture tells me to go back to the one variable analyses: Learn what is to be learned about protein. Learn what is to be learned about fat (perhaps related to bran). The two discussions can be conducted independently, at least until they are better understood.

D = S + N
Data = Signal + Noise

Working up to a linear analysis: Before the beginning, Who, What, Where, Why, When, and How? At the beginning, Stem and Leaf, well-behaved variables. Now I'm ready for a two variable linear analysis. I will start simple by making up some hypothetical data, graphing it, and looking at the graph. Here are my data. Here is my graph, Figure A.

	Data	
Observation	*x*	*y*
#1	10	10
#2	20	14
#3	30	20
#4	40	24

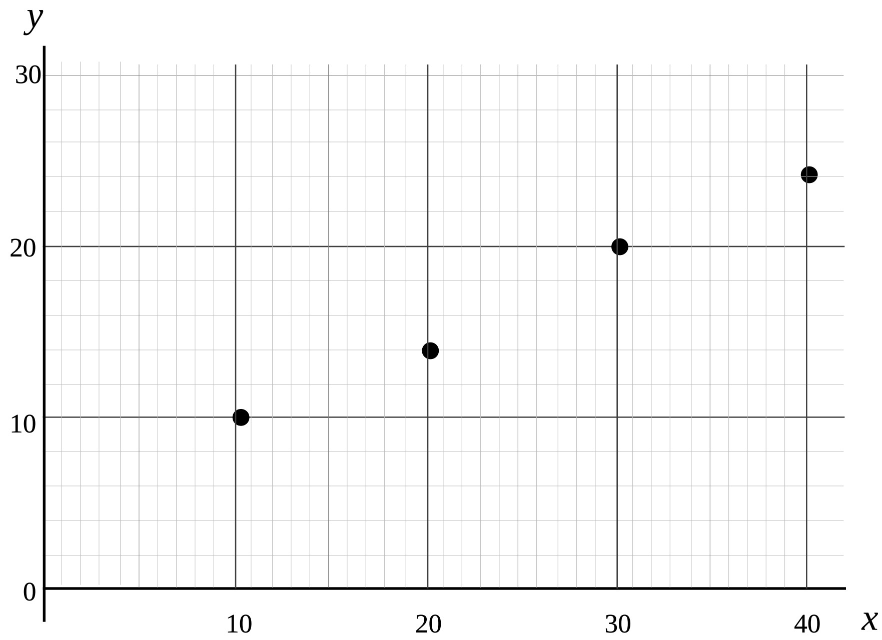


Figure A

Most of what you are ever going to know about these pseudo data you already know — it is apparent in the graph. The relation between x and y is positive, "large x corresponds to large y". The relation looks linear. The relation between x and y has a slope of about one half, with values of y appearing to be directly proportional to the corresponding values of x.

If these were real data, I would probably stop here: Adding the numbers of formalized statistics would be a way of abstracting from reality, summarizing it, and describing its basic regularities. But these are not real data and they allow me to demonstrate the process of adding the formalized numbers. So I will proceed with the obvious in order to show you what is not obvious about the rules for obtaining evidence from such data.

Data = Signal + Noise

Think about a problem in two variable analysis in these terms: Think of the pseudo-equation

$$Data = Signal + Noise \qquad (1)$$

It says that data, the stuff we see and describe has, within it, two parts. One is the signal. That's the message. That's what we are trying to figure out. The other part is noise. Noise may be measurement error, meaningless variation, or a level of complexity which, for the moment, we can not penetrate — so, for lack of understanding it looks like noise.

Now, and this is one of the classical strategies of data analysis, to get at the *signal*, we direct out attention to the *noise*: First we rewrite the pseudo-equation, isolating noise by itself, on the right.

$$Data - Signal = Noise \qquad (2)$$

Then we form a hypothesis about the data and a specific hypothesis about the signal. And then we write a real equation, not a pseudo equation:

$$Data - Hypothesis = Residual \qquad (3)$$

The hypothesis specifies what the data would look like if the hypothesis were correct. And, of course, they do not. Data rarely do. . And so we subtract the hypothesis from the data and see what's left. What's left is called the "residual" and it *should* look like noise. If it doesn't then we reject the hypothesis.

The logic is a bit twisted, I'll admit it. We are interested in the signal; we look at the noise. The reason is that signals do all sorts of interesting things — too interesting, too varied. But noise — we understand noise. We know what noise is supposed to look like. We can recognize noise. When the residuals, equation 3, look like noise, equation 2, it means that when we subtracted our hypothesis about the signal from the data, all that remained was noise: So it was a good hypothesis.

Well-Behaved Noise

This strategy becomes critical for organizing the attack on two variable relations but it is the strategy you have already used to identify well-behaved variables, the same strategy with a different name. With one variable the hypothesis and the signal are pretty rudimentary, hardly deserving of the portentous terms "hypothesis" and "signal". With one variable the "formal" hypothesis is the mean or central value and the implicit hypotheses are embedded in the choice of the unit of measure and the unit of analysis.

With one variable, the pseudo equation

$$Data - Signal = Noise \qquad \text{(Repeating Equation 1)}$$

becomes the real equation

$$Data - Hypothesis = Residual \qquad \text{(Repeating Equation 3)}$$

and very specifically

$$\text{Data - Mean = Residual} \qquad (4)$$

In words, the "residual" is the distribution of data on either side of the mean. If that residual looks like noise: If it is without pattern. If the size of the residuals is small. If the average of this noise is zero and if the distribution of the noise is symmetrical — then the rudimentary one variable hypothesis is correct.

For one variable, in Volume I, this pseudo equation and its interpretation are overkill — an unnecessarily difficult re-statement of the first property of a well-behaved variable. For two variables, this pseudo equation is standard operating procedure.

In simple unsophisticated terms the principle is "Look at the exceptions." You are interested in the pattern, in the signal, but you detect it by looking at the residuals. Returning to the pseudo-equation, you can hypothesize a linear signal,

Data – Signal = Residual.

Hypothesis: Signal is a Linear Relation

But you check the hypothesis by looking at the residuals. Adding it up, there are the three things that a data analyst associates with a line: The intercept, the slope and the residuals. If your residuals look like noise, if the residuals are without pattern, if the residuals are small, and if the mean residual is near zero and the distribution of the residuals is well behaved, then your hypothesis is consistent with the data.

For example, here is standard operating procedure (for lines) in action: Suppose I look at the four points of "data" graphed in Figure A and come up with the hypothesis that these data are approximately constant: The signal is "y is constant." That's wrong: It is a poor hypothesis. But let me pursue it to show how a poor hypothesis leads to poor residuals (residuals that look like they still contain a signal). So continuing, foolishly, I hypothesize that the signal is "y = 5, constant", Figure B.

Because my hypothesis is that "5" is the signal I subtract "5" from the data and look at what's left. (Note, that I have blown up the scale of the graph, to increase the resolution.)

	Data		*Hypothesis*	*Residuals*
Observation	*x*	*y*	$\hat{y}=5$	$y - \hat{y}$
#1	10	10	5	5
#2	20	14	5	9
#3	30	20	5	15
#4	40	24	5	19

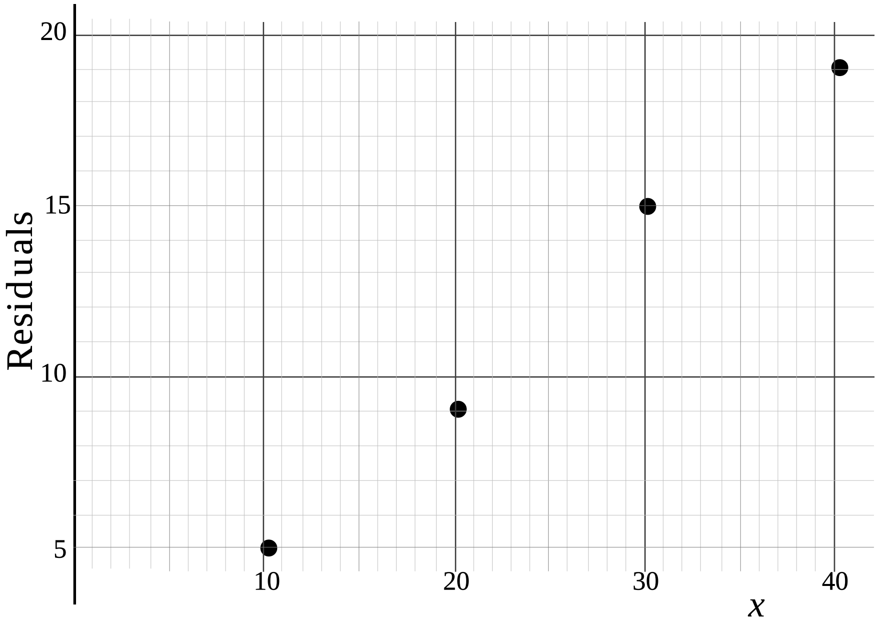


Figure B

Clearly, that's wrong: The residuals show a clear pattern and the residuals are positive, not zero. That is a crude but obvious signal, not noise. So even if the idea is right (which it isn't), the value specified by the hypothesis is wrong. O.K., I can do better. Now I hypothesize: These data are approximately constant with an average value of 17. Now, for the second time, I subtract the hypothetical signal from the data, the signal as it would be if the hypothesis were correct. What do I get (Figure C)?

	Data		*Hypothesis*	*Residuals*
Observation	x	y	$\hat{y}$ =17	y - $\hat{y}$
#1	10	10	17	–7
#2	20	14	17	–3
#3	30	20	17	3
#4	40	24	17	7

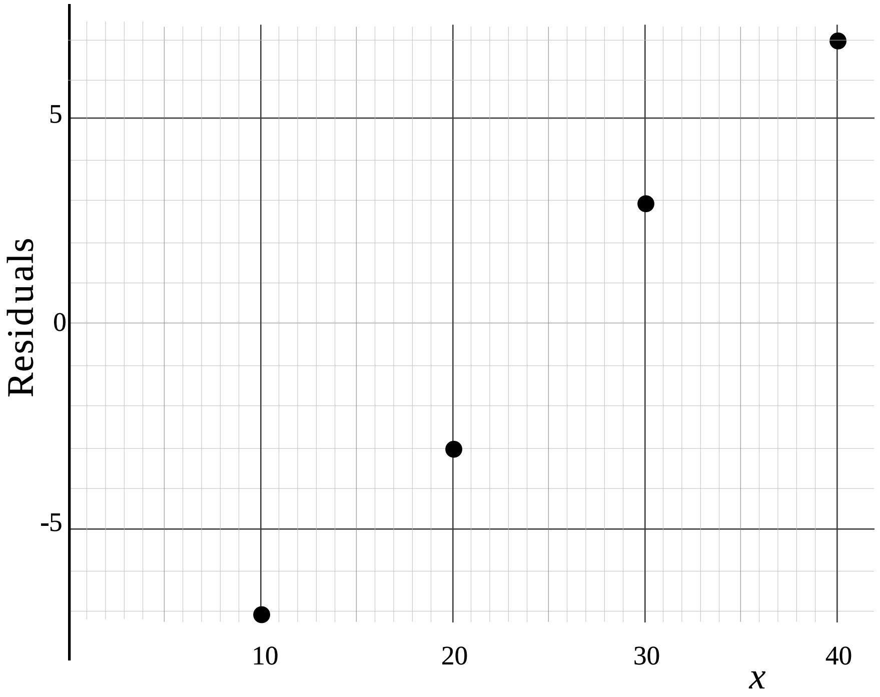


Figure C

Better: The residuals now have an average of zero. But what was omitted from the hypothesis is now painfully obvious in the residuals: The residuals increase quite regularly. So, the data were not constant — back to hypothesis construction.

Now, I'm going to hypothesize that the signal increases with y as a straight line function of *x*: Simply looking at the graph, between x=10 and x=40, I see a run of 30 in the x variable. And over the same interval I see a rise of 14 in the x variable. Strictly by the numbers, that suggests a trial value of "rise/run" = 14/30 = .467 for the slope. I will round that to .5 because I can be relaxed about this first estimate of the slope. I can relax because I would have found slightly different estimates if I had used different points, and because the graph of the residuals will make it apparent that I need to refine the slope — if I need to. So, my new hypothesis is $y = .5x+4$. Testing this hypothesis I subtract the hypothetical signal from the data — I subtract the signal as it would be if my hypothesis were correct — and I get Figure D.

Observation	*Data* *x*	*y*	*Hypothesis* $\hat{y}$ = mx+b b = 4, m = .5	*Residuals* $y - \hat{y}$
#1	10	10	9	1
#2	20	14	14	0
#3	30	20	19	1
#4	40	24	24	0

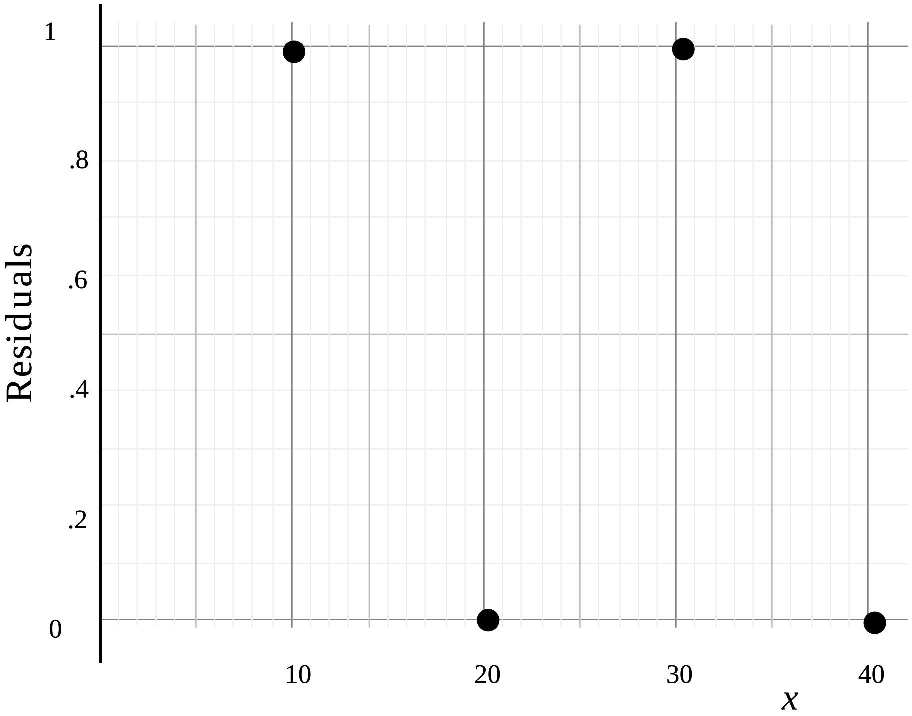


Figure D

That’s more like it: The average of the residuals is small: The magnitudes of the residuals that are not explained bhy this hypothesis are between 0 and 1. And compared to the range of y between 10 and 24, the hypothesis has greatly reduced my uncertainty about y .

Looking closely at what remains in the variation of y, it appears that the residuals are a bit more positive than negative — which suggests that there is a little bit of signal left in these residuals. So I can increase precision by add another 0.5 to the constant (the intercept) in my hypothesis. The residuals show no trace of slope, so I will leave that part of the hypothesis alone. With this added refinement, with the hypothesis y =.5x + 4.5, the linear hypothesis is good, leaving residuals that look like noise (average of zero, no slope), Figure E. It is a good hypothesis.

	Data		*Hypothesis*	*Residuals*
Observation	*x*	*y*	$\hat{y} = mx + b$	$y - \hat{y}$
			b = 4.5, *m* = .5	
#1	10	10	9.5	.5
#2	20	14	14.5	–.5
#3	30	20	19.5	.5
#4	40	24	24.5	–.5

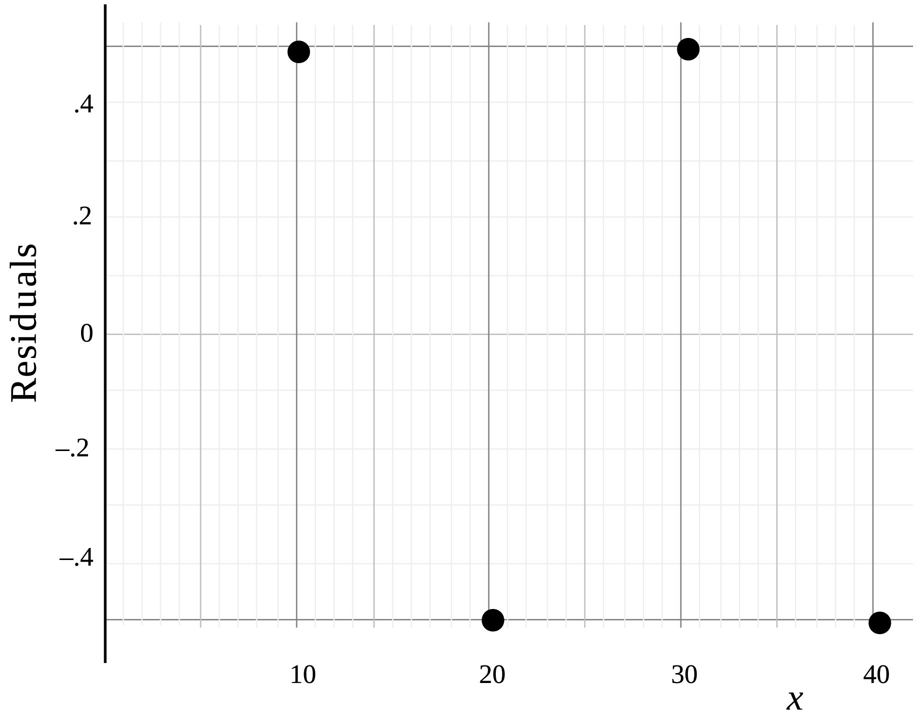


Figure E

Or is it? Is there really no signal left among these residuals? After all of my arithmetic is complete, I am back to the need for human judgment. When I look at what's left, I can't help but observe that the residuals seem to oscillate, up, down, up, and down. Is this a pattern? If there is a pattern in the residuals, then the residuals are not noise and my hypothesis is incomplete. Is this oscillation more signal or is it noise? In truth I can't tell. To build a most-likely answer to that question I would have to place the numbers in context as data: I would have to treat those residuals/noise as a variable, two values are positive, two values are negative and I would have to explore the variable, beginning with the stem and leaf and continuing with the average and variation. I couldn't "prove" anything by such exploration, but then I'm not trying to. I'm looking for ideas. Similarly, if I were not thinking of these residuals as oscillating, up, down, up, down, I might get suspicious of a downward trend among the residuals, noting that the first residual is positive while the last is negative. Is it there? Is this signal or is it noise? Again, I can not tell and no analysis of these four data points will be able to extract an answer from these numbers by mathematical force. Is there a pattern among these residuals? There is certainly no *compelling* evidence of a pattern. But that will not stop me from harboring suspicions and from using those suspicions to hone my questions when I go back to my data source for more exploration.

And are these residuals really small enough to ignore? In truth, I can't answer that question either. Again I would have to place the numbers in context as data: If residuals of this size were errors in the prediction of the domestic products of the United States, year after year, and if the size of the residual corresponded to a one percent deviation between the hypothesis and the data, then I might say "forget it — one percent error in gross domestic product is too small to be taken seriously." But even then I would look for pattern as well as size: If those numbers were gross domestic products for the United States, and the high numbers turned out to be war years, that would give me something to think about — nothing more (no proof, just something to think about — and something to direct my focus at the next stage of my research on GDP), nothing more, but nothing less either.

Doing it with Excel

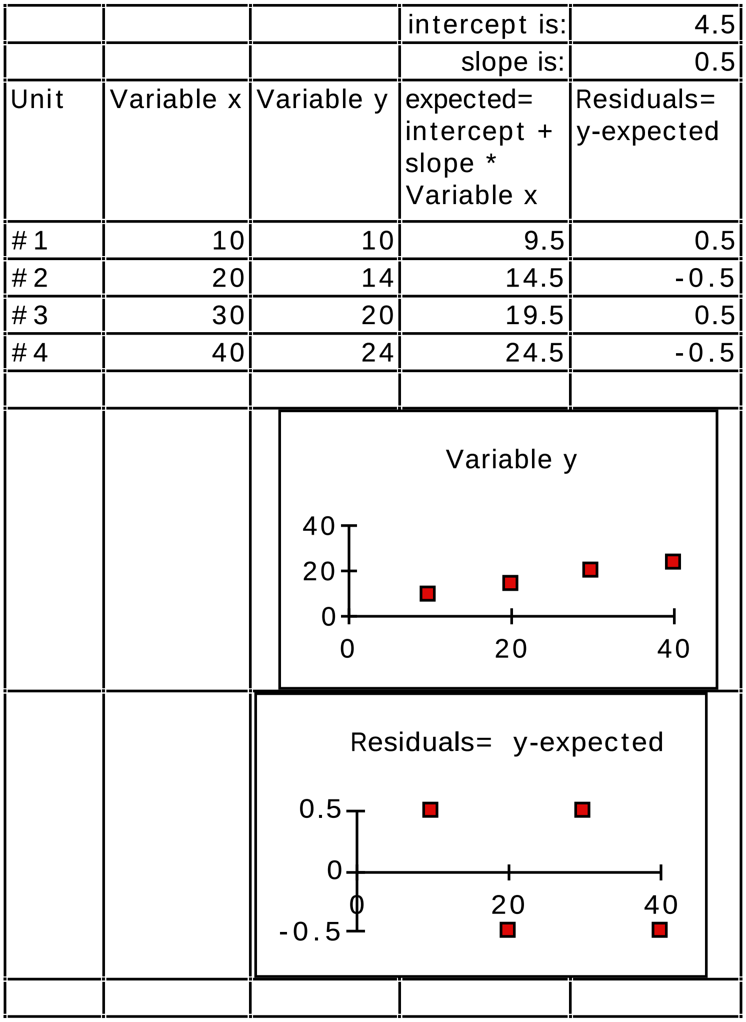
The one thing I do not want you to do with your spreadsheet software on your computer is to have it solve problems for you: Spreadsheet programs are perfectly capable of finding lines for data directly but don't do it. Instead I want you to use the work sheet capacity to carry you through the work step by step — so I know that you know how. (And, because there are choices involved in any data analysis, your program won't get exactly the answer you are looking for.)

So, #1, get your data into the spreadsheet. And be sure to get the units in there among the labels, bags of concrete, 1,000's of people, whatever. Then what? stem and leaf, medians, descriptions, as appropriate for each variable. Each analysis begins at step 0.

Then I'd suggest something like this

			intercept =	Value (You enter this and change it yourself. The formula for the expected value uses this.)	
			slope =	Value (You enter this and change it yourself. The formula for the expected value uses this.)	
Unit	Variable 1 Units	Variable 2 Units	Expected (under the assumptions that the linear hypothesis, with the slope and intercept specified above, is correct	Residual/noise: Observed value of Variable 2 Minus Expected value of Variable 2 under the assumption that the hypothesis is correct	Here you get Excel to do Graph: Across is x, Up is Y. you play around with Excel yo can get it to do somethi visually better: Up is Y ar also up is "Expected" graphing both sequences numbers on one graph
	datum	datum	Formula: = d2 + d1*A3	Formula =B3-C3	Here you get Excel to do anoth graph, lined up below the fir On this graph Across is still But now "up" shows t residuals
	datum	datum	Corresponding to previous		

For example



			intercept is:	4.5
			slope is:	0.5
Unit	Variable x	Variable y	expected= intercept + slope * Variable x	Residuals= y-expected
#1	10	10	9.5	0.5
#2	20	14	14.5	-0.5
#3	30	20	19.5	0.5
#4	40	24	24.5	-0.5

That may take a little bit of doing to set it up. But I want you to set it up and then have Excel do for you what I did in the text: You pick an Intercept, leaving the slope as 0, and then Excel will draw the residuals. You add a non-zero slope, and then Excel will draw the residuals. As you get closer, the residuals should get *less* interesting — that's what noise is supposed to do. You should also be able to keep track of it numerically by watching the average size of the residual go down.

Warning, Excel has a friendly habit of trying to choose a good scale on which to display your graphs. It also has a nasty habit of making a bad choice. Since the scale has a lot to do with what you are going to be able to see, as noted in the text, be careful. If you get something ridiculous in your graph — check the scale. You can intervene, or at least you can intervene on the older version of Excel. I trust I will get advised in class on a variety of different ways to do this. But keep your eye on the purpose of the exercise and choose the scale accordingly.

Beyond Facts

In some of the sciences there is an unstated understanding that facts and theory are separate: According to this idea, numbers are “done” by technicians grappling with the messy stuff of data and reality. Meanwhile, theory is done by theorists, preferably pure theorists, applying their mind to pure ideas and general principles. The idea that such a division works is destructive and dead wrong. It probably grew up as an attempt to make the best of incomplete and divided training among (my generation) of scientists, converting the unhappy fact of division into a virtue: People who work with words are presumed to be theoreticians; people who work with data are presumed to be methodologists.

But verbiage does not make the speaker a theoretician any more than numbers make their user a methodologist. There simply is no such division among active scientists: We construct and develop our ideas by working with the facts; we shape our analysis of the facts in order to shape and test our ideas — there is no division. You will still find a few grand theorists who disdain data. You may ignore them. You will still find a few data crunchers who think that science is a pile of facts, and that greater science is a greater pile. You may ignore them.

The second of the two concepts is a strategy of reverse logic. Non scientists tend to think that scientists “prove” their theories, that we “prove” the laws of physics and “prove” the theory of evolution. We don’t. The trouble is that we can never “prove” anything about the world: The very best we can say is that a theory is consistent with the data — that we have no counter-evidence, not yet. In contrast, disproof can be clear and definitive — just once instance of an event, of something that can’t happen, according to the theory and the theory has been disproved. (or is in need of modification). That’s all it takes. So since disproof can be relatively easy to spot (and proof is impossible) we tend to work by a reverse logic that makes the most of our errors.

And then, bringing the two concepts to together, one of the things we learn the most from is *badly* behaved data — because surprises force us to rethink what’s going on.

The Potato Hypothesis

The place where the ideas and the data merge is in the equations of the hypothesis. Recapitulating, I refereed to the pseudo equation

Data = Signal + Noise

which allowed me to isolate noise as noise = data - signal

Data - Signal = Noise

Then I suggested that the analyst construct a hypotheses about the signal and evaluate the hypothesis by looking at the residuals: if the residuals look like the noise, then the hypotheses captured the signal

Data - Hypothesis = Residual

There is the hypothesis. In the example, the hypothesis was a rather dry statement using a linear equation for y and x as well as estimates of the values for the intercept and the slope. That was about all I could say about a hypothesis when I was using only letters and numbers, x, y, 10, 20, and so forth. But when the hypothesis is about a phenomenon — about some process for which you have gathered the data, then things get much more interesting.

Here are some data. The data are meant to describe the response of crops of potatoes to the application of fertilizer.

The data come from a controlled experiment (Rothamsted Experimental Station Report, 1933) on the effects of increasing amounts of a mixture of the standard crop fertilizers on the yields of potatoes that was carried out in 1933 at the Midland Agricultural College in En gland. the mixture contained 1 part of sulfate of ammonia, 3 parts of superphosphate, and 1 part of sulfate of potash. The amounts were 0, 4, 8, and 12 cwt per acre, the cwt unit, called hundredweight, being actually 112 lb. Owing to natural variability, the yields of potatoes under a given amount of the mixture vary from plot to plot. the yield figures shown for each amount in table 9.2.1 are the means of random samples of four plots.

From *Statistical Methods, Seventh Edition*, Snedecor and Cochran, Iowa State University Press, 1980, pp. 149-150:

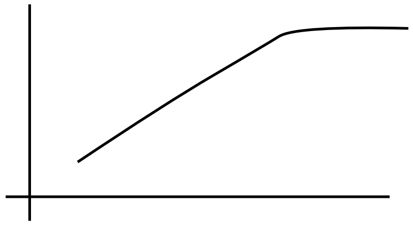
YIELD OF POTATOES IN RESPONSE TO FERTILIZER

Fertilizer in cwt per acre X_i	Potatoes in tonnes per acre Y_i
0	8.34
4	8.89
8	9.16
12	9.50

Now — think! What do I expect and why? That is the half of hypothesis construction that was missing when I used only numbers. What do I expect? Crudely, I expect a positive response to fertilizer, more fertilizer, more potatoes.

Less crudely, I don't know enough about crops to offer a very clever hypothesis. I think of plants as grabbing up nutrients from the soil. The more the nutrient available, the more that will be assimilated by the plant, and the greater the yield — up to a point when the growth approaches the limits of the organism itself.

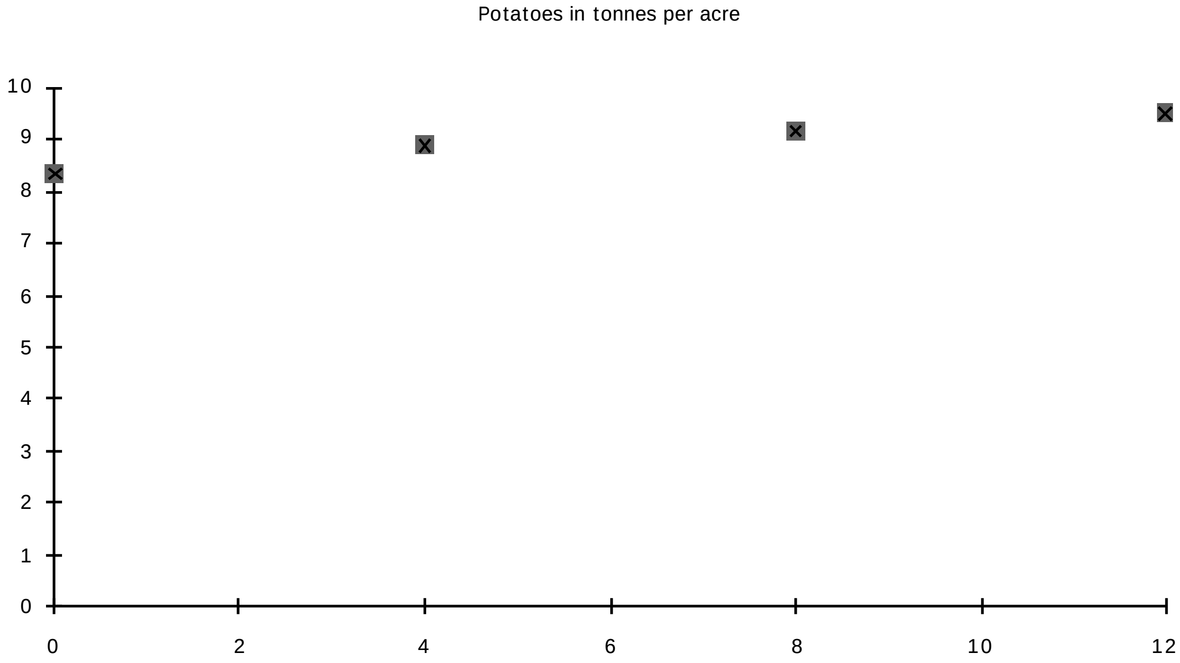
Those are my initial thoughts. Now I crystallize these thoughts into a picture of my expectation, and, if I can, into an equation that will join the thought to the fact. My words suggest a sketch of a linear response, for a part of the range, followed by a bend approaching an asymptote.



What specific equation would I attach to my sketch? I'll have to do it in two parts: Direct proportionality to the fertilizer initially. That would be a line. Later? I would expect something like each additional unit of fertilizer having an equal effect on the distance between the yield and the asymptote. Likely, however, I will not have enough data to track this down very precisely.

Now, I'm ready to look at the data.

			intercept is:	0
			slope is:	0
Unit	fertilizer in cwt per acre	Potatoesin tonnes per acre	expected= intercept + slope * Variable x	Residuals= y-expected
#1	0	8.34	0	8.34
#2	4	8.89	0	8.89
#3	8	9.16	0	9.16
#4	12	9.50	0	9.50



Let me look at this. Before the beginning: Who, what where...? I probably have enough in that descriptive paragraph. Now, are these variables well-behaved? If I evaluate them in terms of their one variable distributions, then I can't tell: The data are from a controlled experiment. That means that I can't really think about the distribution of the applications of fertilizer. The experimenter chose the distribution that, so there is nothing "natural" to be discovered from looking at this distribution to see how it behaved — although it is likely that the experimenter considered the zero, (no fertilizer), and the three applications to be equally spaced. So, I'm unable to consider either the first or the second criterion for a well behaved variable. But the third criterion, linearity, looks useful: The graph suggests that the relation is approximately linear, suggesting that each variable is approximately well behaved. There is more data on the variation of the yield per acre, these are average yields and that average must refer to some distribution. But I don't have it. The fifth criterion is sense. And that is a problem. Whether the data come in pounds, or ounces , or cwt, I can do the arithmetic. But the purpose of the first steps of a data analysis is to feed the intuition of the analyst. So let me tell you about the analyst. I'm an American. My intuition knows pounds and tons, not cwt. What does a cwt of fertilizer look like? Is it a shovel full, or a truckload? I can figure it out. But if I use these units of measure then I am going to slow down my work because every stage will need to be translated until it makes sense. That means I'm not really able to *think* about this. I don't know what 4 cwt of fertilizer is, nor for that matter do I understand a tonne of potatoes. In this primitive sense, I need a change of the unit of measure —so that I can "think" about these without burdening my intuition with translations.

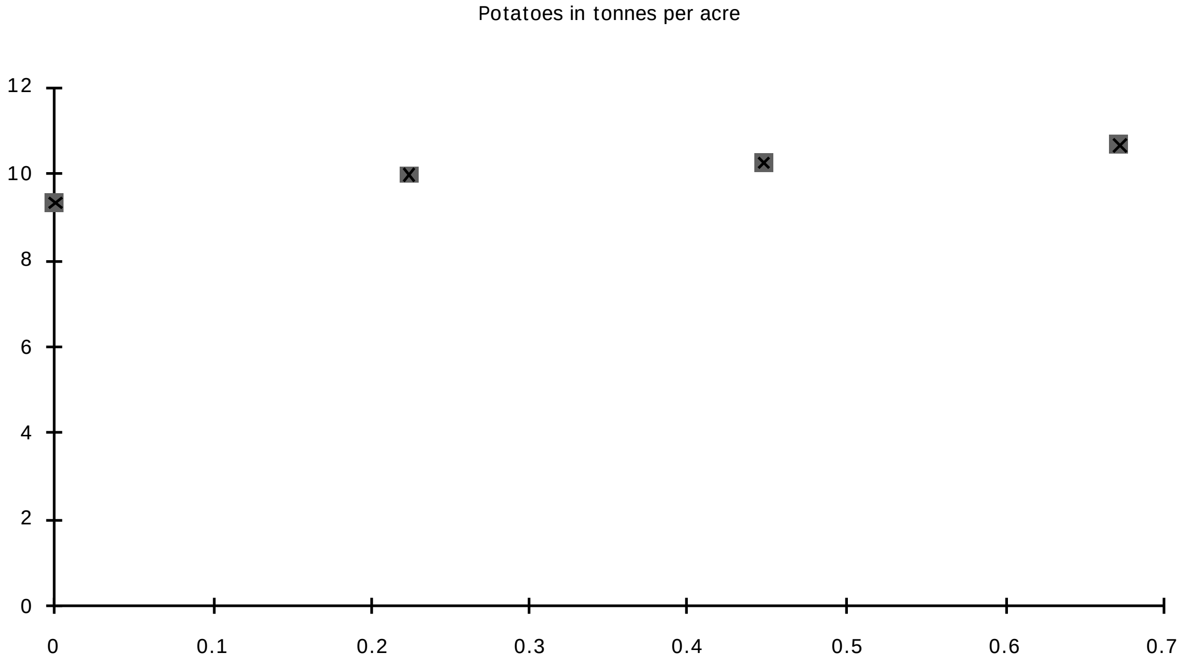
So — to the dictionary. A cwt is a hundredweight. So much for word origins: *cent* (hundred) *weight*, cwt. And what is a hundred weight? It says, "a unit of weight, equal to 100 pounds in the United States and 112 pounds in England". The data are British, and I should take note of that spelling "tonne", also British. So I have 112 pounds of fertilizer for each hundredweight; .056 tons (American) of fertilizer for each hundredweight. Changing the unit measure, the three applications are .22 tons, .45 tons and .67 tons per acre.

My dictionary does not include "tonne", but it has two definitions of "ton", one British. That should be it. A ton, it tells me is "a unit of weight equal to 2,240 pounds avoirdupois (1,016.06 kilograms) commonly used in Great Britain: in full **long ton**, **shipping ton**." Also, "a unit of weight equal to 2,000 pounds avoirdupois (or 907.20

kilograms), commonly used in the United States, Canada, South Africa, etc.: in full **short ton**." So multiplying 8.34 by 2,240 pounds and dividing by 2,000 pounds, the yield of an unfertilized acre is 9.3 tons. And successive yields are 10.0 tons, 10.3 tons, and 10.6 tons per acre.

Converting to these units of measure, I'll start again

			intercept is:	0
			slope is:	0
			intercept is:	0.0
			slope is:	0.0
Unit	fertilizer in tons per acre	Potatoes in tons per acre	expected= intercept + slope * Variable x	Residuals= y-expected
#1	0	9.3408	0.0	
#2	0.224	9.9568	0.0	
#3	0.448	10.2592	0.0	
#4	0.672	10.6400	0.0	



Now, using these units of measure, I'm willing to look. And what do I get from this integration of ideas and number? I get the message that my first ideas were an exhibit of sloppy thinking. I said that yield would be proportional to fertilizer. Note that I drew my sketch pointed toward zero. I was thinking *proportionality* and built it in to my sketch (although my words were more ambiguous). Drawing the line through zero is what "proportionality" means: y is proportional to x, i.e.,

$y = mx$

(not $y - mx + b$)

But that is not the relation shown in the picture. The picture clearly shows that soil is perfectly capable of producing potatoes without the assistance of fertilizer. My thinking was too narrow. I was thinking about fertilizer and yield and human intervention, as if fertilizer were necessary to induce nature to grow crops. And so I simply failed to step back far enough to take into account the very hefty yield that nature can deliver when there is no fertilizer at all.

So, chastised by the data, let me revise my hypothesis by stating it more precisely. I expect the *increase* in yield to be proportional to the increase in *fertilizer*. That idea is matched by a linear equation with an intercept.

I know this revised hypothesis will be in the "ball park" of the data, because I've looked at the data. But I can still test the idea by a straight forward application of reverse logic: Is this hypothesis a fair description of the signal? If it is, then the residuals will look like noise. So, with reverse logic, I direct my attention to the residuals — to evaluate the acceptability of the hypothesis. As for the specific values of slope and intercept, I don't have enough experience to have an expectation or a hypothesis. I'll take these from the data.

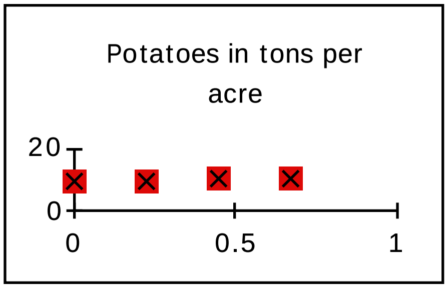
From the data, a first guess for the intercept is obvious, 9.34 tons per acre because that is the observed value of the yield when the fertilizer is 0. A first guess for the slope is also obvious, the vertical rise on the graph is from 9.3 to 10.6. The horizontal run on the graph is from 0 to .67. And the slope is the rise divided by run. That gives me a first guess of approximately 1.93 tons of potatoes per ton of fertilizer.

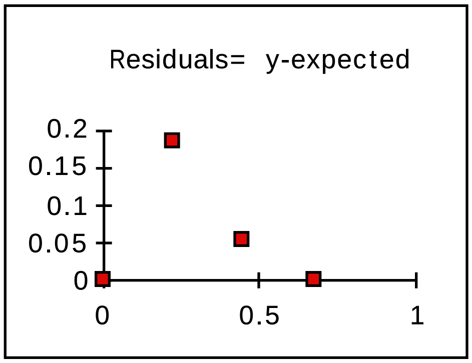
Even without a graph, that is *very* interesting: *1.9 tons of potatoes per ton of fertilizer.* That surprises me: To increase the yield of potatoes by

1.9tons I have to apply *1 ton* of fertilizer to the soil!. (And the 1.9 tons of potatoes are 95% water.) Non agriculturist that I am, non-biologist that I am, I'm astonished, one ton of fertilizer spread on the field to get 2 additional tons of potatoes.

Husbanding my three pieces of information, intercept, slope, and residuals, I'll try placing these estimates for the intercept and the slope into the hypothesis. Then I will look at the pattern of the residuals, asking do these residuals look like noise?

			intercept is:	9.34
			slope is:	1.93
Unit	fertilizer in tons per acre	Potatoes in tons per acre	expected= intercept + slope * Variable x	Residuals= y-expected
#1	0	9.3408	9.340	0.0008
#2	0.224	9.9568	9.772	0.18448
#3	0.448	10.2592	10.205	0.05456
#4	0.672	10.6400	10.637	0.00304





Residuals with respect to the line with
intercept = 9.34, slope = 1.93.

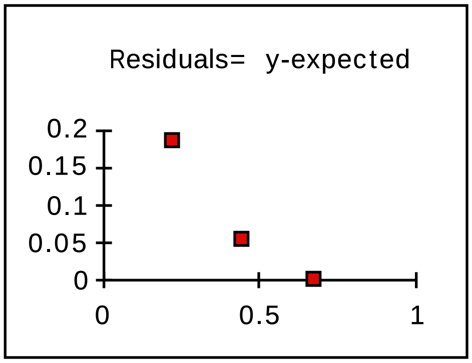
No, these residuals certainly suggest a pattern among the residuals, not noise. So, by reverse logic, my hypothesis doesn't represent the signal. My equation was wrong. And what is important is not just that my equation was wrong. What is important is that my thinking was wrong. Writing my thoughts into the equation allowed me to test the ideas with data. My ideas were wrong again.

I'll keep going with these data but from here forward I am not going to be able to test my subsequent . I'll call the next idea a hypothesis. But by now I've seen so much of the data that I am really just writing a hypothesis to fit the facts — to test it I would need other data.

With that proviso, I *think* that what I see in these data is that the soil alone (without fertilizer) was capable of a yield of approximately 9.3 tonnes per acre. The first addition of 0.2 tons of fertilizer to the untreated soil added about 0.6 tons to the yield, .2 tons in, .6 tons out. And each additional .2 tons produced an increase of .33 tons, .2 tons in .3 tons out. In words, the untreated soil, without fertilizer, comes within 88% of the maximum yield achieved with the heaviest application of fertilizer (calculating 9.34/10.64 = 88%). The first application of fertilizer to the untreated soil has a disproportionately large gain as compared to subsequent applications. Perhaps the effect of the fertilizer is catalytic as well as directly nutritive so that the initial application enables the organism to use nutrients that were already present. This is a one-shot effect. As a result further increments of fertilizer will add only the nutritional effect.

For my graph, the only thing to be "tested" after all this handling of the data is that the last two increments are approximately equal. Limiting my graph to the last two increments (the last three data points):

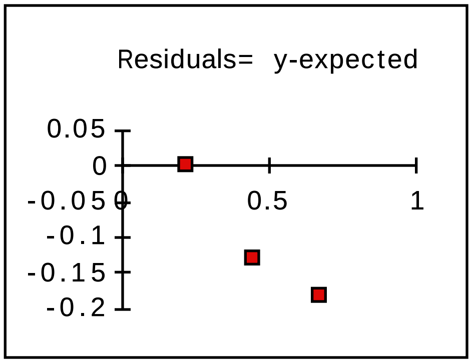
			intercept is:	9.34
			slope is:	1.93
Unit	fertilizer in tons per acre	Potatoesin tons per acre	expected= intercept + slope * Variable x	Residuals= y-expected
#1				
#2	0.224	9.9568	9.772	0.18448
#3	0.448	10.2592	10.205	0.05456
#4	0.672	10.64	10.637	0.00304



Residuals with respect to the line with intercept = 9.34, slope = 0.097.

From that, I can bring the first residual down to zero by treating its value, .184, as a positive signal. I remove this positive signal from the residuals by adding it to the hypothesis.

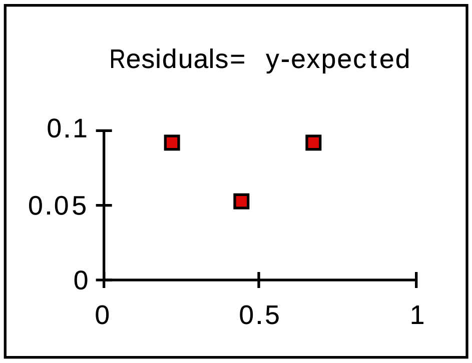
			intercept is:	9.524
			slope is:	1.93
Unit	fertilizer in tons per acre	Potatoesin tons per acre	expected= intercept + slope * Variable x	Residuals= y-expected
#1				
#2	0.224	9.9568	9.956	0.00048
#3	0.448	10.2592	10.389	-0.12944
#4	0.672	10.64	10.821	-0.18096



Residuals with respect to the line with intercept = 9.524, slope = 1.93.

Then I can work on the slope suggested by the residuals: The residuals show a vertical descent of -.181 over a horizontal run of .448. Calculating -.181/.448 = -.405, that is a signal of -.405 in the residuals. I take it out of the residuals by adding negative -.405 to the hypothesis.

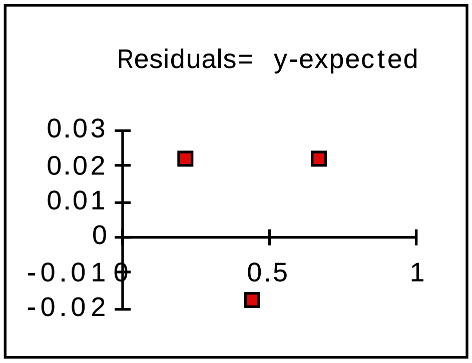
			intercept is:	9.524
			slope is:	1.525
Unit	fertilizer in tons per acre	Potatoes in tons per acre	expected= intercept + slope * Variable x	Residuals= y-expected
#1				
#2	0.224	9.9568	9.866	0.0912
#3	0.448	10.2592	10.207	0.052
#4	0.672	10.64	10.549	0.0912



Residuals with respect to the line with intercept = 9.524, slope = 1.525.

In turn, I can go back to the intercept for fine tuning. Note that it *is* fine tuning, because the size of these residuals is small. It now shows an intercept of about .07 in the residuals. I take this .07 out of the residuals by adding it to the hypothesis, getting:

			intercept is:	9.594
			slope is:	1.525
Unit	fertilizer in tons per acre	Potatoesin tons per acre	expected= intercept + slope * Variable x	Residuals= y-expected
#1				
#2	0.224	9.9568	9.936	0.0212
#3	0.448	10.2592	10.277	-0.018
#4	0.672	10.64	10.619	0.0212



Residuals with respect to the line with intercept = 9.594, slope = 1.525.

Those are about as harmless a set of residuals as I can imagine but, again, these residuals aren't really a test of anything because, now that I am down to three data points, if there is any error at all (which there always is), this is what it has to look like, down/up or up/down. (Anything else, for example, up/up would have been interpreted as a slope and then removed from the residuals.)

So, what do I know? Let's put it in order, noting that the most obvious things were obvious, once there was a picture.

I know that the untreated fields averaged ____ tons of potatoes per acre. (I know that by converting the data to common units and reading the data.)

I know that the lightest application of fertilizer received the greatest return in terms of additional tons of potatoes per ton of fertilizer, about 2.75 tons of potatoes per ton of fertilizer. (I know the effect from looking at the residuals. I know the number by simple calculation directly from the data.)

I know that additional fertilizer had a lower marginal return of potatoes about 1.525 tons of potatoes per son of fertilizer. (I know this by examination of secondary residuals and by fitting a line to the last three of the four data points.)

I also know that my initial ideas had very little to do with what was found. Certainly it is not true that growth is proportional to fertilizer. It is not even true that additional growth is proportional to additional fertilizer. Instead additional growth is realized at one rate for the first application of fertilizer and at a lower rate for greater applications of fertilizer. And, finally, these data provide no evidence of a diminishing rate of return.

In case you hadn't noticed, there is an almost inverse relation between the utility of each piece of information gained from various aspects of these data and the amount of work that was necessary to extract that information. Most of what was learned: By learning that my hypothesis was sloppy (and wrong), and then by getting a description of the actual behavior of the data — these things can be read directly from the data and the second graph. The less obvious things, uncovered with page after page of technical virtuosity, added detail: Before I went through this detailed procedure I knew that the slope for the last three data points was about 1.5 (in the phrase ".2 tons in, .3 tons out"). Now I know it is more like 1.525 than 1.5 — which is to say the largest technical display in this analysis was attached to the smallest gain of real information, increasing the precision of the estimate from approximately 1.5 to approximately 1.525.

That is a hard blow to the ego of the data analyst. By the end of an analysis you are focused on your most technically sophisticated efforts. But very often, the technological sophistication added little to what was known when you graphed two well-behaved variables, one against the other. That was your real act of sophistication: The real act of sophistication is getting the initial display right so that the most important results become the most visible features of the graph. Once the initial display is under control, the obvious results became obvious.

The next blow to the analyst comes in the write up. Truth is, nobody cares how hard you worked. “Heh! I’ve got 14 pages of work here (in manuscript).” Tough. That’s how much work it takes, but no one cares. They want to know, in this case, about fertilizer and potatoes. Ontogony may recapitulate phyologeny (in biology), but in data analysis, the report does not recapitulate the research — except to leave some evidence for the cognoscenti: The cognoscenti (other analysts) need assurance that you actually did the research and they need sufficient information to allow them to reproduce the work for themselves if they care to.

Data from the Rothamsted research station present a picture of the increases in crop yield that may be obtained through the application of fertilizer. The data show that the lightest application of fertilizer brought a return of 3 tons of potatoes per ton of fertilizer, approximately a seven percent increase compared to the untreated fields. Additional applications of fertilizer at rates that were double and tripe the initial application achieved a smaller increase of approximately 1.5 additional tons of potatoes per additional ton of fertilizer

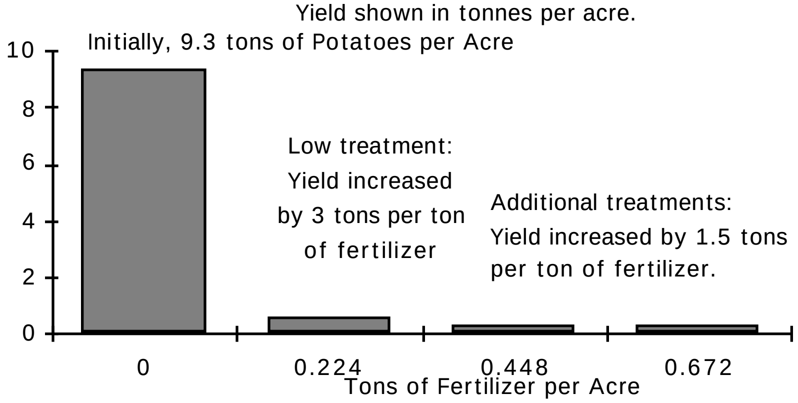
The data come from a controlled experiment (*Rothamsted Experimental Station Report, 1933*) on the effects of increasing amounts of a mixture of the standard crop fertilizers on the yields of potatoes that was carried out in 1933 at the Midland Agricultural College in England. The fertilizer contained 1 part of sulfate of ammonia, 3 parts of superphosphate, and 1 part of sulfate of potash, with applications and average crop yields as shown in the table below.

Yields of Potatoes Corresponding to Different Amounts of Fertilizer

Fertilizer *(cwt per acre)*	Average Yield *(tonnes per acre)*
0	8.34
4	8.89
8	9.16
12	9.50

Secondary source: *Statistical Methods*, Seventh Edition, Snedecor and Cochran, Iowa State University Press, 1980, p 152.

Initial and Incremental yield of Potatoes in response to Fertilizer



Exercise: From *Statistical Methods*, Snedecor and Cochran, Iowa State University Press, 1980, p. 153:

Two problems:

In a controlled experiment on the effects of increasing amounts of mixed fertilizers on sugar beets conducted at Redbourne, Lincs, England, in 1933, the mean yields of sugar beet roots and tops (n=5) for each amount X are as follows:

X (cwt/acre)	0	4	8	12	16
Roots (T/acre)	14.42	15.31	15.62	15.94	15.76
Tops (T/acre)	7.48	8.65	9.74	11.00	11.65

Problem 1: Analyze the data for roots.
Problem 2: Analyze the data for tops.

In a controlled experiment on the effects of increasing amounts of mixed fertilizers on sugar beets conducted at Redbourne, Lincs, England, in 1933 the mean yields of sugar beet roots and tops for each amount X are as follows

X (cwt/acre)	0	4	8	12	16
Roots (T/acre)	14.41	15.31	15.62	15.94	15.76
Tops (T/acre)	7.48	8.65	9.74	11.00	11.65

Analyze and describe the yield in Roots as a function of fertilizer.

From Snedecor, p. 153. (Original source not fully specified.)

The Soybean Hypothesis: Growth of an Organism v/s Growth of a Population

For a second example, let me escalate from four data points to seven — same basic numerical technique, but different data and, therefore, different hypotheses, and a different deployment of the numerical technique in pursuit of the hypotheses.

Again from Snedecor and Cochran:

From *Statistical Methods, Seventh Edition*, Snedecor and Cochran, Iowa State University Press, 1980, p 152:

> Example 9.2.1-- The following are measurement of heights of soybean plants in a field — a different random selection each week (Wentz, J. b. and Steward, R.T., *J. Am. Soc. Agron* 16 (1924):534.

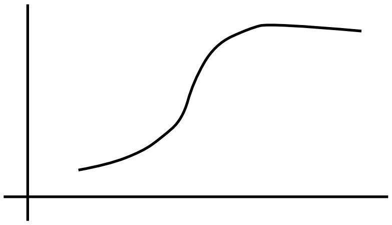
Age. X (wk)	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Height, Y (cm)	**5**	**13**	**16**	**23**	**33**	**38**	**40**

Before the beginning, "Who, What, Where, ...?" As usual these preliminary questions must co-exist with pedagogical duplicity, by which I mean it is necessary to act *as if* I were subjecting these data to the same close scrutiny that I would exercise were I doing research whose quality and outcome depended on the quality of the data. The textbook is excellent. The data are often classics of the trade.

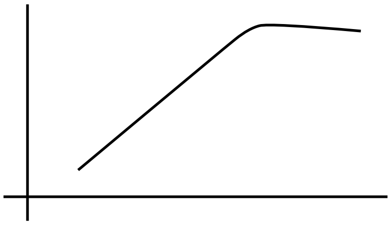
Examining the data I am a little wary after my previous experience: When data come in a series, as did the fertilizer data, look for trouble at one end or the other: Boundary points are more likely to be "different" than middle points. Here, the initial boundary (age 0, height ?) is just missing.

O.K., let me think about growth. My reflexive expectation about growth is that things grow exponentially: Little things grow slowly, big things grow rapidly — both little things and big things grow in proportion to their existing size: A tiny thing can not sustain a rapid growth. But then, at the other end, there are limits as growth approaches the limit of the resources or the capacity of the organism.

So I expect accelerating growth at the beginning, decelerating growth near the end. The accelerating growth should be approximately exponential and, therefore, it should be linear when the unit of analysis is the logarithm of the height. But then — on second thought I'm not so sure: What is it that increases? Is it the height of the plant, as given? Or is it the weight of the plant, the "biomass" . If it is the latter then I should be looking at the cube root of the height (the relation between volume and extension where form is constant). But then the log of the cube root of the height the result would be linear, or non-linear, as the log of the height itself is linear or non-linear. So, I am still looking for something like a linear relation between the logarithm of the height and the age of the plant.



Pattern in Centimeters

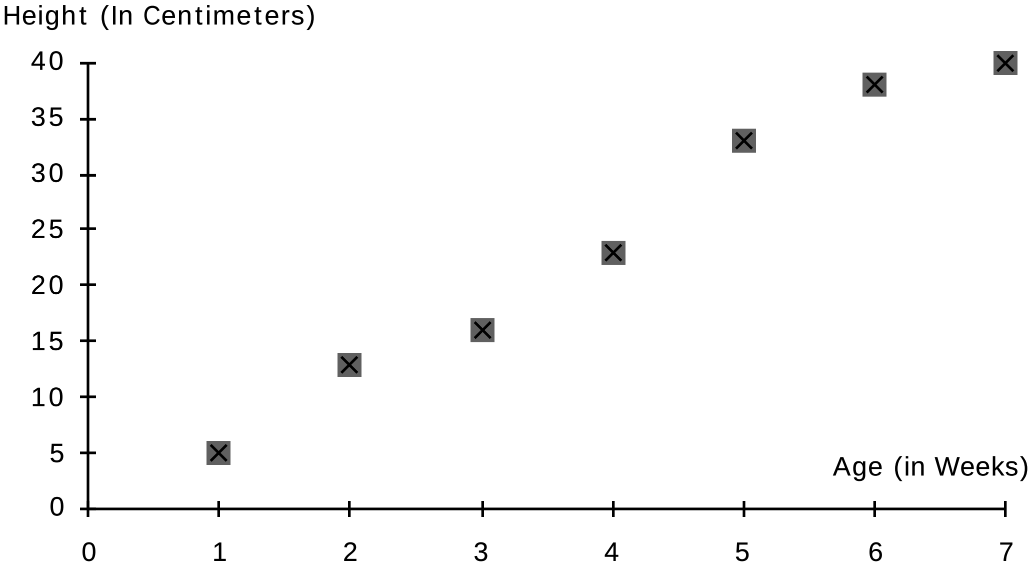


Pattern in Logs

As for one-variable analysis, once again this is a controlled experiment. That implies that values of age were controlled by the experimenter and there is nothing to be learned about nature from the examination of the numbers for age. But I will be able to use a two-variable criterion, looking for linearity (in the log of the height).

Hypothesis in hand, I am committed. Let me see what the data have to teach me.

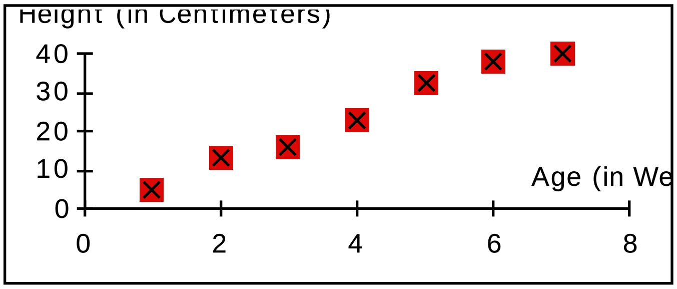
			intercept is:	
			slope is:	
Unit	Age of Soybean (In Weeks)	Mean Height of Soybean Plants (in centimeters)	expected= intercept + slope * Variable x	Residuals= y-expected
#1	1	5		
#2	2	13		
#3	3	16		
#4	4	23		
#5	5	33		
#6	6	38		
#7	7	40		

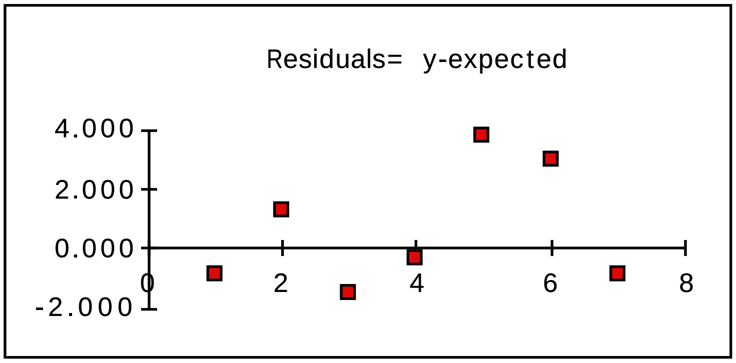


This is discouraging. It looks nearly linear — although it will take a look at the residuals to test that appearance. But my thinking led me to expect an increasing rate of growth as the plant became large enough to support more growth. My thinking is wrong. That is the way to use data: It is very good at disproving hypotheses while it can never really prove a hypothesis. So we set up the data analysis in order to get the maximum mileage from what data analysis is good at — rejecting hypotheses. This hypothesis, and the line of reasoning behind it, is wrong.

Is it really linear? Let me look at the residuals. I will try an intercept of 0, 0 weeks, 0 centimeters. I will try a slope of 6.33 centimeters per week, calculating a rise of 35 centimeters, and a run of 6 weeks, 35/6 = 5.83

			intercept is:	0
			slope is:	5.833333
Unit	Age of Soybean (In Weeks)	Mean Height of Soybean Plants (in centimeters)	expected= intercept + slope * Variable x	Residuals= y-expected
#1	1	5	5.833	-0.833
#2	2	13	11.667	1.333
#3	3	16	17.500	-1.500
#4	4	23	23.333	-0.333
#5	5	33	29.167	3.833
#6	6	38	35.000	3.000
#7	7	40	40.833	-0.833





Confirmed, there is little evidence of an exponential curve departing from my linear hypothesis. Confirmed, my hypothesis is not supported by the data. It appears that there is a remaining slope in these residuals. The original picture suggests that the boundary point, 7 weeks, may be different — which fits at least a little of my hypothesis. If I were to remove that last point from the residuals, there would be a definite upward slope. If I am to leave it in, then the residuals suggest a little increase of the variance of the residuals toward the end. I am going to ignore that last data point to see what parameters I get for the best line.

The most important characteristic of residuals is their pattern, not their size. Here the “pattern’ is an absence of an upward curve. That being accepted, I can attend to the size of the residuals, apart from their pattern. That makes the “fitting” process easier. Here I will compute the squared residuals and the mean of the squared residuals, and then select a slope and an intercept for six of the seven data points that gives me a good fit in the sense of least squares.

Starting with the first estimate of the intercept and the slope

intercept is: 0
slope is: 5.833333

Unit	Age of Soybean (In Weeks)	Mean Height of Soybean Plants (in centimeters)	expected= intercept + slope * Variable x	Residuals= y-expected	Squared Residuals
#1	1	5	5.833	-0.833	0.69444444
#2	2	13	11.667	1.333	1.77777778
#3	3	16	17.500	-1.500	2.25
#4	4	23	23.333	-0.333	0.11111111
#5	5	33	29.167	3.833	14.6944444
#6	6	38	35.000	3.000	9
#7	7	40	40.833	xxxx	xxxx
				Mean Residual	Mean Squared Residual
				0.916667	4.75462963

The mean residuals (without squares) are slightly positive on the average. Likely then I can improve the fit by transferring the average from the residuals to the hypothesis.

intercept is: 0.92
slope is: 5.833333

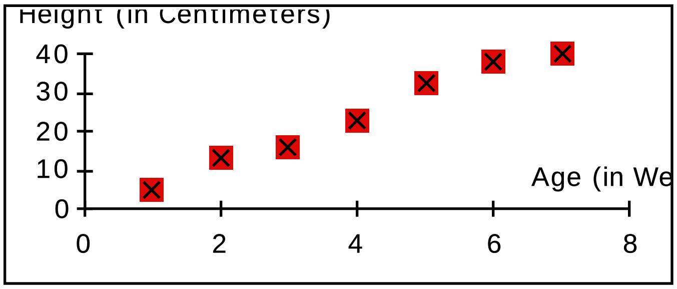
Unit	Age of Soybean (In Weeks)	Mean Height of Soybean Plants (in centimeters)	expected= intercept + slope * Variable x	Residuals= y-expected	Squared Residuals
#1	1	5	6.753	-1.753	3.07417778
#2	2	13	12.587	0.413	0.17084444
#3	3	16	18.420	-2.420	5.8564
#4	4	23	24.253	-1.253	1.57084444
#5	5	33	30.087	2.913	8.48751111
#6	6	38	35.920	2.080	4.3264
#7	7	40	41.753	xxxx	xxxx
				Mean Residual	Mean Squared Residual
				-0.00333	3.91436296

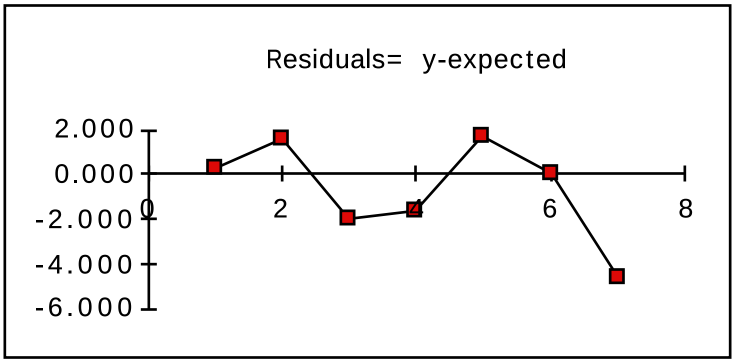
Yes, that reduces the squared residual by about 15 per cent, from 4.75 to 3.91. With more work, my best is:

intercept is:	-1.86667
slope is:	6.628571

Unit	Age of Soybean (In Weeks)	Mean Height of Soybean Plants (in centimeters)	expected= intercept + slope * Variable x	Residuals= y-expected	Squared Residuals
#1	1	5	4.762	0.238	0.05668934
#2	2	13	11.390	1.610	2.59056689
#3	3	16	18.019	-2.019	4.07655329
#4	4	23	24.648	-1.648	2.71464853
#5	5	33	31.276	1.724	2.97151927
#6	6	38	37.905	0.095	0.00907029
#7	7	40	44.533	-4.533	xxxx

Mean Residual	Mean Squared Residual
3.4E-15	2.07
	Square Root of the Mean Squared Residual
	1.44





That gets rid of about half of the squared residuals. To represent this by a number, think of an analogy to the mean, the variance, and the standard deviation: In the analogy, the mean is the number closest to the data in the sense of least squares. The variance reports the mean of those (least) squares. And the standard deviation reports an average error converted (by the square root) to a unit of measure that matches the unit of measure of the variable. Here, by analogy, the square root of the squared deviation reports the size of the residuals as a standard error, which is, in this case, 1.4 centimeters (computed for data points one through six).

Looking at these residuals, there is no simple pattern among the residuals. (Allowing that the human eye always sees *something*, I can imagine an inverted “W” here. But I don’t take it seriously: Failing any *simple* connection between such a pattern and the substance of the data I will describe the residuals as patternless.)

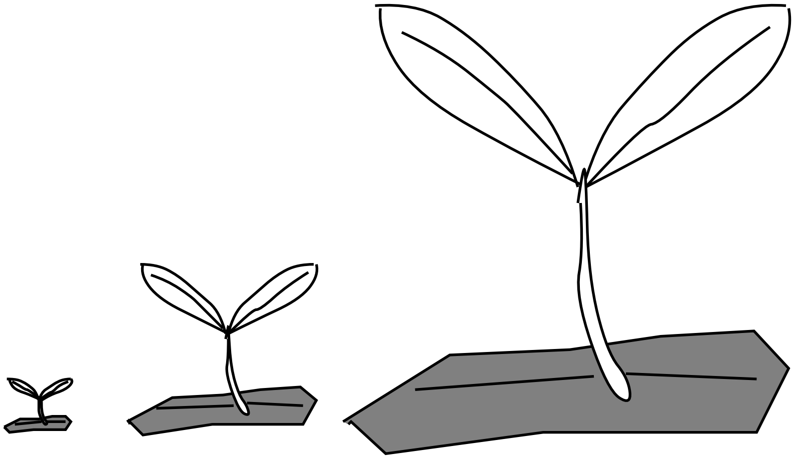
The last data point was not used in this computation because it looked suspicious. The graph supports the exceptional nature of this last point by comparing the observed value to the hypothetical value extrapolated from the six preceding points. Although this is only one data point, it suggests a declining rate of growth for the seven week old plant.

The slope suggests a growth of 6.6 centimeters per week, including a hypothetical week 0, when the bean would have been 2 centimeters below the surface of the ground (the intercept).

The data certainly look simple, the data are linear to a reasonable approximation, for the data preceding the last observation. But this is a negative result: Remember I hypothesized an exponential growth, and I told you why. I was wrong. So the numbers of this linear description are not the end of my work. On the contrary, they tell me I have some explaining to do. There is something wrong with my thinking that needs to be discovered, uprooted, and replaced.

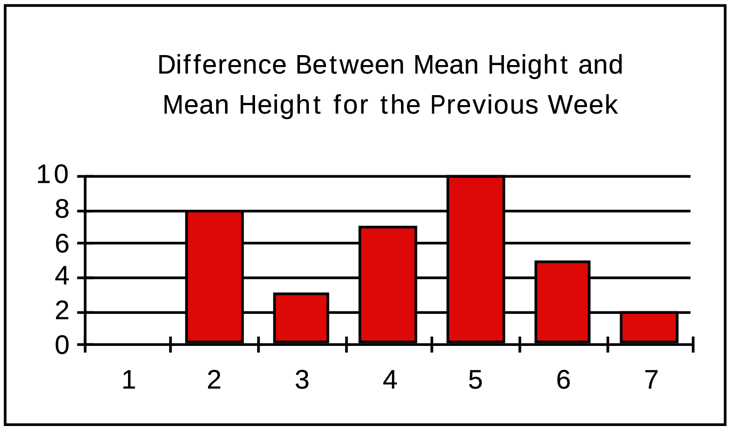
My research process for this dilemma is straight forward — a phone call to my nearest biologist. He advises me that I am thinking like a sociologist. Sociologists look at populations that increase and reasonably well you might expect the rate of growth of a population to be proportional to the number of potential parents. Sociologists look at wealth, which tends to grow in proportion to investment.

Wrong model: Plants don't grow that way. They don't get born looking like a plant, grow up looking like the same plant, but larger, and mature looking like the same plant, but larger.



Wrong model. Plants don't grow that way (nor for that matter do people, although people do, on the average, tend to maintain the same number of arms and legs and heads as they grow older.) Plants like soybeans and trees (and long bones of people), grow at their growing tips. These specialized cells appear to do their work at an approximately constant rate (for a while). The growth of single organisms obeys different laws as compared to the growth of populations.

So — the work is almost complete. But not yet. My friend's story hasn't really explained it. His is a new hypothesis. And you know what to do with a new hypothesis: Test it. No fair doing all that work to test the old hypothesis, and destroy the old hypothesis — and then fail to subject the new explanation to equal scrutiny. The new hypothesis says "constant growth" — at least for a while. So let me do a simple graph of the amount of growth, each week.



This is very confusing. I look at this and I see pattern: Infancy (of a soybean) large initial growth. Then youth, steady growth increasing to a spurt of growth (week 5 has three times the growth of week 3), finished off with a period of declining growth.

This is a problem: I have a story here. I like it. It seems to fit the data. But how did this regularity escape me earlier? What I was trying to do here was create a picture that would be a compelling exhibit of constant growth (if the growth were constant).

But actually I had already constructed a picture of residuals from constant growth — those were the residuals from the straight line (and a straight line is a model of constant growth). What is the difference? There are two differences. The first major difference was my mind set: When I looked at residuals from the straight line I was looking for a rising curve that would encourage me to go forward to test the initial hypothesis (exponential growth). I didn't find it. My mind was set to evaluate the first hypothesis. I did and I rejected it. The second major difference is that in this graph I reflexively left out the first week because the first week could not be compared to an earlier week — it is different.

Now with the new hypothesis I have a different mind set. I'm evaluating this new hypothesis and ready to be critical of the new hypothesis. And because I am focusing on differences, I reflexively use less data — I only examined the data for which there was a difference to be examined. As a result, subjecting the new hypothesis to the same degree of skepticism as I applied to the original hypothesis, I *suspect* that I see a pattern (where previously I saw an inverted "W" which I took to be meaningless.

So what is the *truth* about soybeans? What *is* going on here? I have lots of stories now. What is going on? This is embarrassing — seven data points are keeping me fully occupied. So, I need to think: I leave my desk, watch a movie, go for a run, stare at the ceiling — all the time thinking about soybeans.

By now I've looked at these seven data points so carefully that it should be no wonder that I can come up with a story that will fit the data — but that doesn't mean that the new story is wrong. Here is what I've learned from these data:

This is a report that will never become a finished report on its own. For me, or for a research group, it is an internal memo — a step in the research. It definitely ends my first hypothesis. But does it allow me to say that growth follows 2 or 3 phases: Rapid initial growth, then a longer period of initially slow but increasing growth, climaxed by a peak, followed by a decline? If I were to pursue that description solely on the numbers, I would need more numbers. For example, if I had daily data, 49 data points, and the pattern persisted, then I would know that the ups and downs were real. 7 data points can do a lot of bouncing. 49 data points following the same broad trends are more credible.

But that's a numerical fix that avoids getting into the real research, out to the soybean fields, in to the biology of plants. What I would really do, were this 1924 (when the work was published), or were this Spring and I had a few soybeans, is to *look* at the plants: If I withdraw from the numbers and my computer I do know a little bit about plants. They start with one or two seed leaves that are basically collapsed but present in the seed — ready to catch some sun very quickly. Seeds germinate, doing nothing visible for a few days, then pulling these cotyledons from the seed cover and out into the sun. What is the germination period for a soybean plant? I don’t know. I would watch and look for parallels between the “stages” of growth and the phases I suspect I see in the data: Growth from the seed in the first week or two. Then I watch the plant and watch the data: Is there a pause as the plant switches over to a new structure (as compared to the seed growth). Is there a spurt followed in a few weeks by the end of new growth and change to favor other functions — reproduction, or root growth. I would look at the phasing in my data and attempt to match it to structural changes in the plant itself. For the moment, these seven data points, and the techniques of data analysis have done their job which is to lead me (much more deeply than I had intended) into questions about the physiology of soybeans.

Data from Wentz and Stuart (*J Am Soc. Agron* 16, 1924) p. 534) demonstrate the pattern of growth in the soybean plant. Starting from a seed beneath the soil, the plants achieved an average length of 40 centimeters 7 weeks later. While the length of the average plant increased by an average of 5.7 centimeters per week, the change per week is irregular, showing a maximum of 10 centimeters growth in the fifth week which is three times larger than growth in the third week but almost matched by growth in the second week.

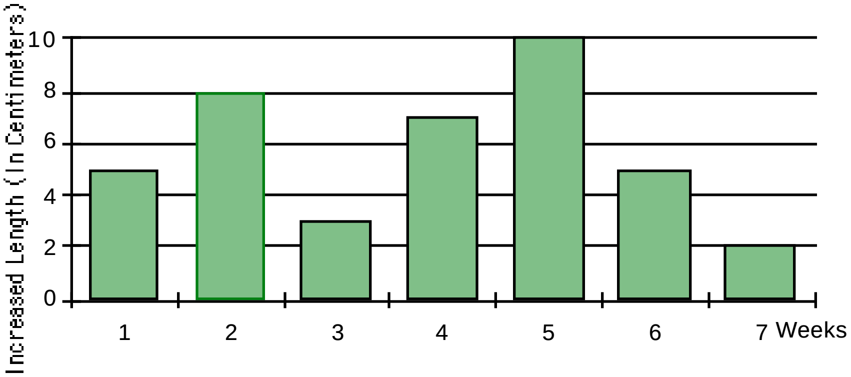
Further research should look for physical changes in the plant which may explain these differences in the growth rates. For example: What is the germination period? At what time are the seed leaves followed by the growth of stem and new leaves. When does the plant cease to grow? When does it flower? These stages in the growth of the plant may explain its initial growth, 5 centimeters in the first week, 8 centimeters of additional growth in the second week, plummeting to 3 centimeters growth in the third week. After the third week there is an gradual rise, with maximum growth in the fifth week, diminishing to the slowest growth exhibited at any time in the seventh week of the plants' lives.

Average Height of Soybean Plants Compared to the Age of the Plant

Age *(weeks)*	Length *(centimeters)*
1	5
2	13
3	16
4	23
5	33
6	38
7	40

Secondary source: *Statistical Methods*, Seventh Edition, Snedecor and Cochran, Iowa State University Press, 1980, p 152.

Initial Growth and Weekly Increase (In Centimeters)



From Snedecor, Statistical Methods, page 153

In a controlled experiment on the effects of increasing amounts of mixed fertilizers on sugar beets conducted at Redbourne, Lincs, England, in 1933, the mean yields of sugar beet roots and tops (=5) for each amount X are as follows

X (cwt/acre)	0	4	8	12	16
Roots (T/acre)	14.42	15.31	15.62	15.94	15.76
Tops (T/acre)	7.48	8.65	9.74	11.00	11.65

Describe the response of the root crop of sugar beets to applications of fertilizer.

Height and height

From Snedecor, op. cit., p. 175.

Stature (inches) of Brother and Sister
(illustration taken from Pearson and Lee's sample of 1401 families)

Family Number	1	2	3	4	5	6	7	8	9	10	11
Brother	71	68	66	67	70	71	70	73	72	65	66
Sister	69	64	65	63	65	62	65	64	66	59	62

Original: Pearson, K. and Lees, A. *Biometrika* 2 (1902-3):357.

Big(ger) Data Sets

Height and Weight

I think that a reasonable person has to be stunned by the amount of detail that can be found in as few as 4 data points (fertilizer and potatoes data), or as few as 7 data points (soybean growth data). It surprises me, and it surprises me again every time it happens. Nevertheless, the world often presents us with data involving many data points, considerably more than 8, or 80, or 800. There is no simple answer about what you do because the effect of large data sets usually can be counted on to produce two opposite effects on the self assurance and equanimity of the data analyst. Large data sets can fill in the blanks where you think you see a pattern, but need more data to be sure that the overall pattern is obeyed: In the soybean growth data it is possible that the inverted "W" is nothing but noise. If daily data traced the lines of the "W", conforming to it and filling in the spaces, it would be reassuring that the pattern (whatever its meaning) is not noise. In this regard data sets can be re-assuring — when the details agree with the overview I have more confidence that the overview is "real".

Large data sets can also have the opposite effect, confirming that patterns that could not possibly be correct, are correct. And then leaving us with the daunting task of explaining them.

For example take a very close look at the relation between heights and weights of adult (5,000 adult British women), as shown in Table __. I consider it obvious that taller people will tend to be heavier: Certainly there will be tall but thin people who weigh less than short but heavy people. But there will also be tall heavy people and short thin people so that, on the average, taller people will tend to be heavier.

This supposedly obvious statement must stand or fall by the facts. "Obvious" or not an assertion about fact has to be tested. And unfortunately, there are some very strong exceptions in these data. Is it true that taller women are heavier? On the average, yes, but there are some curious exceptions even "on the average". For example, what are the odds that a 5'2" woman will weigh 140 pounds as compared to 134 pounds? Answer: The odds are just about even. At 5'2" there are 95 woman weighing in at 134 pounds, 101 weighing in at 140.5 pounds.

Now , for comparison, what are the odds that a women of 5’4” will weigh 140 pounds as compared to 134 pounds? These woman are 2 inches taller. Presumably they are likely to be heavier than the first set of women at 5’2”. So the odds favoring 140 pounds as compared to 134 pounds should be greater. Obvious. And false: Among the shorter woman there are about as many woman at the heavier weight as at the lighter weight. By comparison, among these taller woman there are 25% *fewer* woman at the heavier weight than at the lighter weight, 138 versus 175.

How do I explain that? Well, first off, I hope I can explain it away as a quirk, as something weird, as something that doesn’t need to be explained. But I can’t get away with that one. There are at least eight instances of this effect in these data and the frequencies involved are very large and, therefore, among the most believable numbers in the data.

						Column Sums						Totals
	5	33	254	813	1340	1454	750	275	56	11	4	4995
278.5 lbs						1						1
272.5 lbs												0
266.5 lbs						1						1
260.5 lbs							1					1
254.5 lbs												0
248.5 lbs					1	1						2
242.5 lbs							1					1
236.5 lbs							1					1
230.5 lbs					2				1			3
224.5 lbs					1	2	1					4
218.5 lbs			1		2	1		1				5
212.5 lbs				2	1	6		1	1			11
206.5 lbs				2	2	3	2		1			10
200.5 lbs			4	2	6	2						14
194.5 lbs				1	3	7	7	4	1			23
188.5 lbs			1	5	14	8	12	3	1	2		46
182.5 lbs			1	7	12	26	9	5		1	2	63
176.5 lbs			5	8	18	21	15	11	7		2	87
170.5 lbs			2	11	17	44	21	13	3	1		112
164.5 lbs		1	3	12	35	48	30	15	5	3		152
158.5 lbs			8	17	52	42	36	21	9			185
152.5 lbs		1	7	30	81	71	58	21	2	2		273
146.5 lbs		2	13	36	76	91	82	36	8	1		345
140.5 lbs		1	6	55	101	138	89	50	8			448
134.5 lbs			15	64	95	175	122	45	5			521
128.5 lbs		1	19	73	155	207	101	25	3			584
122.5 lbs		3	34	91	168	200	81	12	1	1		591
116.5 lbs		3	24	108	184	184	50	8				561
110.5 lbs		5	33	119	165	124	22	4				472
104.5 lbs	1	3	33	87	95	35	6					260
98.5 lbs	2	5	29	59	45	16	3					159
92.5 lbs		6	10	21	9							46
86.5 lbs		1	5	3								9
80.5 lbs	2	1	1									4
Weight												
	54in	56in	58in	60in	62in	64in	66in	68in	70in	72in	74in	Height

Reproduced from Kendall and Stuart, *op. cit.,* p. 300.

Figure 6.1

Distribution of Height and Weight for 4,995 Women, Great Britain, 1951.

Weight	58in	60in	62in	64in	66in	68in
152.5 lbs			81	71	58	21
146.5 lbs			76	91	82	36
140.5 lbs			**101**	**138**	89	50
134.5 lbs		64	**95**	**175**		
128.5 lbs		73	155			
122.5 lbs	34	91				
116.5 lbs	24	108				

Reproduced from Kendall and Stuart, *op. cit.*, p. 300.

Figure 6.1
Anomalous Subsets of the Data for the Distribution
of Height and Weight, Great Britain, 1951.

So, how do I explain these anomalous facts of height and weight? For now, I do not explain it. (See Levine, 1993, Chapter 6, for at attempt at an explanation.) The reason I present it, without presenting the methodological trick that makes everything clear, is to make the point that data analysts work at different levels of detail for different purposes. Very often, data analysts are working toward a fairly modest degree of description: Does an increase in x correspond to an increase in y? When possible, we will put a number on it: For a unit increase in x what is the average increase in y? In that context the word "model" is applied to the straight line, y = mx + b.

Sometimes we want to get into the detail of the data because we want to get into the detail of the mechanism. Then the word "model" is applied to a theory of the mechanism at work behind the data. Most of the time we are not prepared to work with that level of commitment to the work of explaining the mechanism behind the data.

So, satisfying myself with simple description, what is the approximate relation between height and weight:

y = mx+b,

$$y \text{ pounds} = m\left(\text{pounds}\big/_{\text{inch}}\right) x \text{ inches} + b \text{ pounds}$$

Even with this modest, though fairly typical goal, there is no line that is going to “fit” these data: We know that immediately. Among the woman at 5’2”, for example, the weights range from 92.5 pounds (9 women) to 248.5 pounds (1 woman). The equation y = mx+b is simply not capable or predicting 92.5 pounds for one woman at 5’2” and 248.5 pounds for another woman at 5’2”.

What we try to do is to predict the mean. For women at 5’2”, the mean is 130.22 pounds. For women at 5’4”, the mean is 134.59 pounds. We ask for the “line of means”: the best line for predicting the means, y = mx+b? (It would be entirely valid to predict medians. Probably because of the historical difficulty of numerical computation without computers, the line of means became the most often used procedure and survives as the more common procedure.)

In a later chapter I will introduce formulas that make it easy to work with the entire data set when computing the line of means, known also as the “regression line”. But for now, begin with the mean weight at each height and describe the relation between (mean) weight and height:

Homework:

DATA

Height (inches)	Mean weight (pounds)
54	92.50
56	111.41
58	122.05
60	124.43
62	130.22
64	134.59
66	140.48
68	146.37
70	157.32
72	163.41
74	179.50

The Relation Between Height and Weight[1][2]

A report on the heights and weights of approximately 5,000 British women published in 1951, indicated that at that time women who were five feet tall weighed 125 pounds, on the average. The relation between height and weight indicated taller women weighed more than shorter women at an average of approximately 3.4 pounds per inch.[3]

The detail of the data, graphed in Figure 1, show the average weights at each height and a reference line sketching the approximately linear relation.

The linear relation is a reasonably good predictor of the average weight: The median error, predicting the average, is approximately five pounds — although the average error for predicting the weight of particular women, instead of the average, may be presumed to have been much greater.

However, even with these small errors with respect to the average, the pattern of the errors, shown in Figure 2 indicates a distinct non-linearity. The errors follow a distinctly "S-shaped" pattern, observed weight is lower than predicted among the shortest women, observed weight is higher than predicted among the tallest women, and observed weight the doubly-bent "S-pattern" in the middle. Were we to exclude both ends of the data, then the average pounds per inch would be estimated at approximately 2.6 pounds per inch. While restricting the range of variation, excluding the very short and the very tall, would improve estimates and change the ratio, to about 2.6 pounds per inch, nevertheless it would not describe the phenomenon because even within this restricted range the deviations show a distinctly non-linear pattern, shown in Figure 3.

While it is clear that there is a non-linearity to these data it is clear that single-bend transformations, like the cube root, would be incapable of explaining the pattern of these deviations. It is also true that the end points of the data are based on relatively few women, 5 women at 54 inches, 4 women at 74 inches, but the regularity, albeit an unexplained regularity of the deviations suggests that the deviations are not the result of random error due to the small number of observations.

[1] Excel spread sheet attached.

[2] Note: If this seems to bear a distinct resemblance to the previous write-up, there's a reason. Presentation takes time, always more than I expect. The first one took a considerable amount of time, but the second one began by pasting the new data and the new graphs into the old write-up, and then changing what needed to be changed.

[3] The data are reproduced in detail in Kendall's The Advanced Theory of Statistics, Volume II, pp. 300 and 319. The detailed data show not only the average weight for each height but the detailed numbers of women counted at weight ranging from 80.5 pounds to 278.5 pounds.

DATA		COMPUTED VALUES		
Height (inches)	Mean weight (pounds)	Expected Values (in centimeters) (Expected values under the hypothesis that a 60 inch woman weighs 124 pound and that weights deviate from 124 pounds at the rate of 3.4 pounds per inch.)	Error (in centimeters) (Error defined as yield minus expected yield.)	Number of Women
54	92.5	103.6	-11	5
56	111.41	110.4	1.01	33
58	122.05	117.2	4.85	254
60	124.43	124	0.43	813
62	130.22	130.8	-0.6	1,340
64	134.59	137.6	-3	1,454
66	140.48	144.4	-3.9	750
68	146.37	151.2	-4.8	275
70	157.32	158	-0.7	56
72	163.41	164.8	-1.4	11
74	179.5	171.6	7.9	4

Average magnitude (absolute value) of error in centimeters: 3,609 pounds

Table 1
Heights and Average Weights for 4,995 British Women

Source: Reproduced from Kendall's The Advanced Theory of Statistics, Volume II, pp. 319, reproduced, in turn, from Women’s Measurements and Sizes, London, H.M.S.P., 1957.

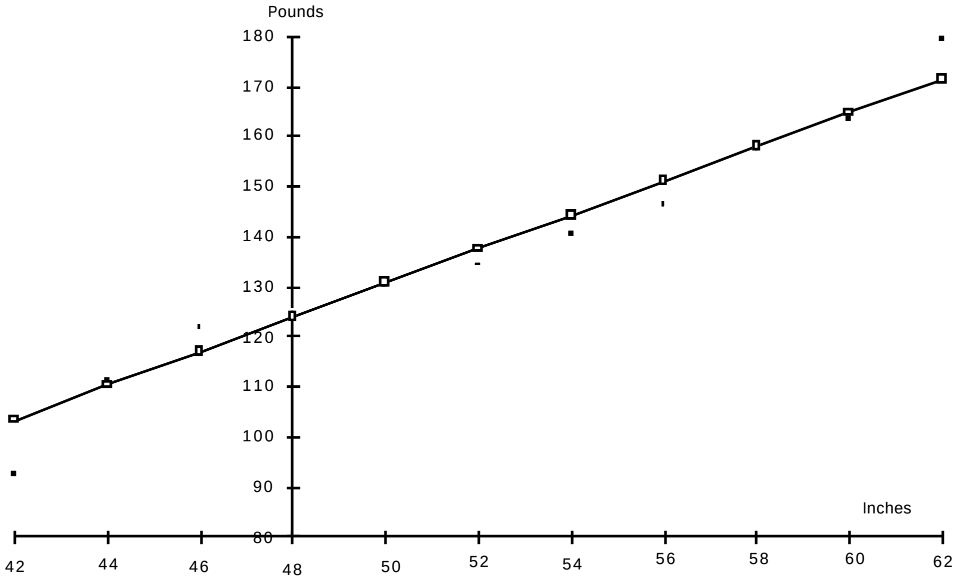


Figure 1
Height and Average Weight for 1495 British Women, circa 1951.

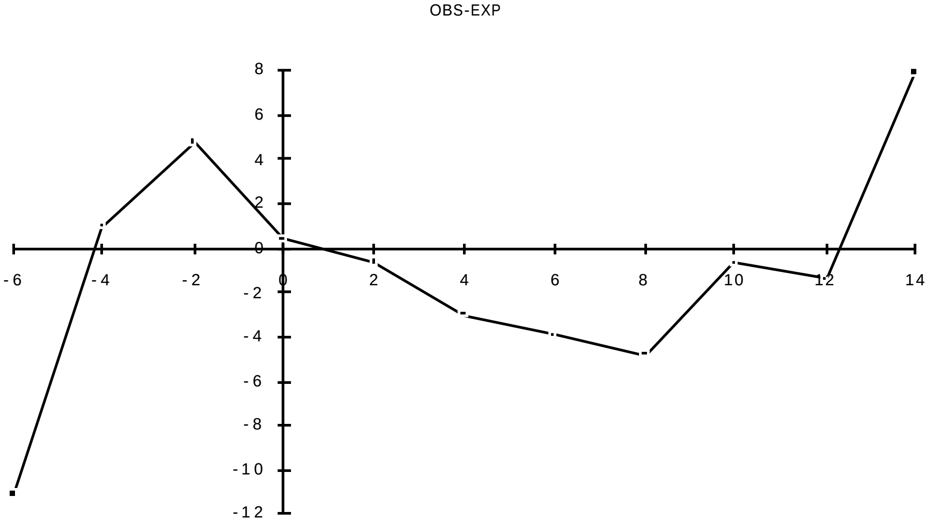


Figure 2
Errors of Prediction, Comparing Observed Weight to Predicted Weight

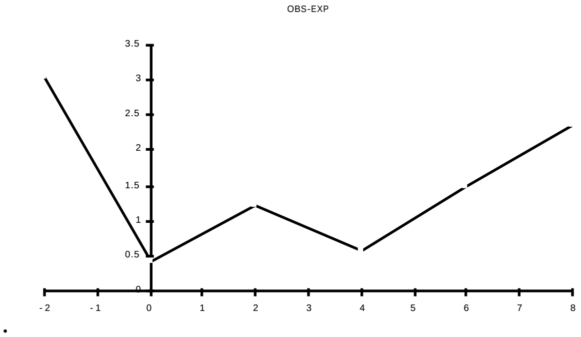


Figure 3
Errors of Prediction, Within the Restricted Middle Range of Heights Demonstrating Persistent Non-Linearity.

Log Lines
(Changes, of Logs)

Logs: Once again, as with averages and lines, you know the math. of what we are about to discuss. *Using* the math is another matter and that is my subject.

What is a logarithm? There are several ways of defining it that come to the same thing. Let's take the secondary school approach (as compared to the calculus/college approach) because it is probably the useful one for data analysis. The story of logarithms begins with the story of exponents. We know for example that in some cases the multiplication of two numbers in exponential form can be seen as addition, addition applied to the exponents. Thus, if one number is 2^3, two cubed, and another is $2^{4,}$ two to the fourth power, then the multiplication of these two numbers can be expressed, equally well as multiplication, multiplication of the numbers themselves, or as addition, addition of their exponents. That is = 8

	Exponential Form		Standard Form
First number:	2^2	=	4
Second number:	2^3	=	8
Third number:	2^5	=	32

Product, written as a multiplication problem:

$$4 * 8 = 32$$

Product, written as addition problem (adding exponents):

$$(2^2)*(2^3) = 2^{2+3} = 2^5 = 32$$

Or, in symbols

$b^p*b^q = b^{p+q}$

That is a quick review of "high school" mathematics of exponents, with perhaps one more useful fact that should be remembered: What you have, above, is that multiplication, applied to the original numbers, can (as above) be re-expressed as addition, applied to the exponents. This basic correspondence between multiplication of simple numbers and addition of exponents implies another correspondence between exponentiation, of the simple numbers and multiplication of the exponents

Operation applied to the simple numbers	Operation applied to the exponents	
Multiplication ——————>	Addition	$b^p*b^q \longrightarrow b^{p+q}$
Exponentiation ——————>	Multiplication	$(b^p)^r \longrightarrow (b^{rp})$

These properties of exponents depend on an implicit assumption, specifically, that the "thing" being exponentiated stays the same from number to number: Above, when I worked with numbers, the "thing" was always 2, I always used exponents of 2. Here, symbolically, the thing is b, I have used exponents of "b". It doesn't matter what number I exponentiate as long as I am consistent. The number I use is called the base. and its exponent is the logarithm of x base b.

$$b^{log_b x} = x$$

You are accustomed to seeing these things in two forms: In logs base 10, called “common logs”, and particularly in the sciences and mathematics in logs base “e”, where it is useful to use a special constant e, approximately equal to 2.71828. A third relatively rarely-used base is simply 2 (often used in information theory and related areas of computer science). Thus

Number	Log Base 10 “Common Logs”	Log Base e “Natural Logs”	Log Base 2
1	.0	.0	.0
1.01	.0043	**.01**	.014355
1.02	.0086	**.02**	.028569
1.03	.0128	**.03**	.042644
2	.3010	.69	**1.**
4	.6021	1.39	**2.**
8	.9031	2.08	**3.**
10	**1.**	2.30	3.321928
100	**2.**	4.61	6.643856
1000	**3.**	6.91	9.965784

Before we use these things in data analysis, let me dwell a little on the actual values. Why, for example, would you prefer one base to another? Mathematically, it makes no difference. Mathematically, these are just three different ways of doing the same thing — three ways of taking advantage of the properties of exponents. But if not mathematically, then practically it makes a difference for the usual reason: Appropriate numbers serve to direct attention to regularities in the data. Note, for example, that each of these bases for the logarithm gives you easily remembered numbers in some range, less easily remembered numbers in other ranges. So, if you are using data in which binary arithmetic is important (as for computers), use logs base 2: It takes log2(x) bits (rounded up), to record the value of an integer.

If you are using data in which items vary by orders of magnitude (i.e., by powers of 10), then use logs base 10: In 1990 the population of the United States was approximately 250 million (250,000,000), or “two times ten to the eighth”. (Counting digits to the left of the decimal point minus one.) This is sometimes referred to as “scientific notation”, “2.5 x 10^8” — where you recognize “8” as the logarithm, base 10 (approximately). t The population of Canada was “two times ten to the seventh”. Ah, I know immediately that the United States population is an order of magnitude larger than the population of Canada. Base 10 helps. Using the raw number, “26,538,000”, it is mentally clumsy to compare it to other numbers. Remember it as approximately 10^7 and you know immediately that it was an order of magnitude smaller than the size of the United States.

Base e, is very convenient when you are working with small percent increases, as in economics and in population data for which an economy, or a bank account, or a population is likely to grow at a rate of between 1 and 10 percent per year. It is convenient because of a transparent correspondence between ratios and their logarithms when the ratios are in this range. In this range I have need for log tables, I can do it “in my head”. For ratios close to 1

	Ln of the Ratio	
Ratio	(Logarithm, base e) (Two digits)	(Four Digits)

1	0	0
1.01	0.01	0.0100
1.02	0.02	0.0198
1.03	0.03	0.0296
1.04	0.04	0.0392
1.05	0.05	0.0488
1.06	0.06	0.0583
1.07	0.07	0.0677
1.08	0.08	0.0770
1.09	0.09	0.0862
1.10	0.10	0.0953
1.11	0.10	0.1044
1.12	0.11	0.1133
1.13	0.12	0.1222
1.14	0.13	0.1310

Within this range, computing logarithms, base e, is easy. In turn this makes it simple to work w compound interest on interest rates within this range. For example, let me do some compound interest "in m head" asking how long it takes for the principle to double at various rates of interest. The rough answer what accountants call "The rule of 70." Which means take the interest rate, divide 70 by the interesting ra That is the doubling time. Or, you can ask the question, if I know the doubling time, then what is the intere The logic is straightforward:

Suppose I start with a principal of $100. At one percent growth per year, how long will it take to doub At the end of one year I will have $100*1.01. At the end of two years I will have $100*1.01*1.01. How ma years in a row do I have to apply the multiplier 1.01 before I get $200?

$100*1.01*1.01*1.01* ……… •1.01 = $200

If I use logs base e, and if commit to memory that the log of 2 is approximately .70, then the "trick" is convert this multiplication problem to an addition problem:

In logs, How many years in a row do I have to add .01 before it adds up to ln 2.

ln(100) + ln(1.01) + ln(1.01) + ln(1.01) ………+ln(1.01) = ln(100) + ln(2) combining terms

$\ln(100) + n(\ln(1.01)) = \ln(100) + \ln(2)$

This is the logarithmic equivalent of asking how many multiplications by 1.01 (how many additions of ln 1.01) do I need to achieve doubling. How do I solve it? I remove the troublesome log(100), no need for it

$n(\ln(1.01)) = \ln(2)$

I pluck the logarithms from my memory. Substituting the values of the logarithms into the equation:

$n(0.01) = .70$

And now I know:

$n = 70$ (approximately)

At 1% per year, the principal will take approximately 70 years to double. How long will it take for my money to double at 2% per annum? In detail:

$n(\ln(1.02)) = \ln(2)$

Inserting the logarithms, base 3:

$n(0.02) = .70$

And therefore, approximately,

$n = 35$ (approximately)

At 2% per year, the principal will double in approximately 35 years.

How many years will it take for my money to double at 10% per annum? In detail:

$n(\ln(1.10)) = \ln(2)$

Inserting the logarithms, from memory:

n(0.10) = .70

And now I know

n = 7 (approximately)

At 10% per year, the principal will double in approximately 7 years.

When would I use such things, other than to impress my bank by doing compound interest in my head? Frequently. For example, we are about to analyze the rate of growth of the population of the United States. In two hundred years, from 1790 to 1990, it grew from approximately 3 million people ???? to approximately 250 million people ????. So: What was the average annual rate of increase? As usual, I can do a good approximation to the full data analysis quickly, and in my head:

Here’s the logic: I focus on doubling. How many times has the U.S. population doubled? That’s, 3 million, 6 million, 12 million, 24 million, 48 million, 96 million, 192 million, 384 million: It doubled a little more than 6 times in 200 years. So, roughly, it doubled every 33 years. Ah: If it doubled in 33 years, then the average annual rate must have been approximately 2%.

I’ve barely begun to analyze these data, but I’ve got a baseline. The population increase was approximately 2% per year.

I put it to you:

In 1945 the U. S. federal government took in $45 billion dollars in revenue. In 1992 the federal revenu
trillion dollars. What was the average annual rate of increase?

In 1945 the U.S. federal outlays for national defense were 82 billion dollars. In 1992 the outlay was
What was the average annual rate of increase?

In 1945 the U.S. federal outlays for human resources were 2 billion dollars. In 1992 the outlay was 777 billion. What was the average annual rate of increase? (Source, Statistical Abstract of the United States, 1992, Table 491, page 315.)

In 1960 U. S. national health expenditures added up to 27.1 billion dollars. In 1990 the total was 666 billion. What was the average annual rate of increase?

In 1960 U. S. national health expenditures added up to 143 dollars per capita. In 1990 the figure was 2,566 dollars per capita. What was the average annual rate of increase? (Source, same. Table 135, page 97)

In 1959-60 U. S. personal income per capita was approximately $2,200. In 1990 the figure was 18,720 dollars . What was the average annual rate of increase? (Source, same. Table 678, page 431)

The Slide Rule

For about 300 years most scientists had an intuitive grasp of logarithms as an indirect consequence of using something called the slide rule. So, as a curiosity and, to aid your intuition, let's discuss the slide rule.

I assume that the original reason for their use was the computational simplicity they introduced for multiplication — when you haven't yet invented the computer you care very much about such computational simplicity. I'm guessing, but I'd make a small bet that this device goes back to at least the fifteenth century, Regiomantanus, and stayed in very heavy use until at least the 1970's. Today there is no need for them But we do need the intuitive comfort with logarithms that the slide rule created. Assuming that they are not even manufactured any more, offer you as a paper cut out slide rule — serviceable but somewhat fragile.

The slide rule is a machine that adds logarithms physically by adding-up physical lengths. Starting with the obvious, start with 4 times 8. Multiplying 4 times 8 is simple. Multiplying 4 times 8 using logarithms, you add log 4 to log 8 and get an answer which is equal to log 32.

On the slide rule you do the same thing by adding the length corresponding to 4 to the length corresponding to 8 and reading that the result is the length corresponding to 32:

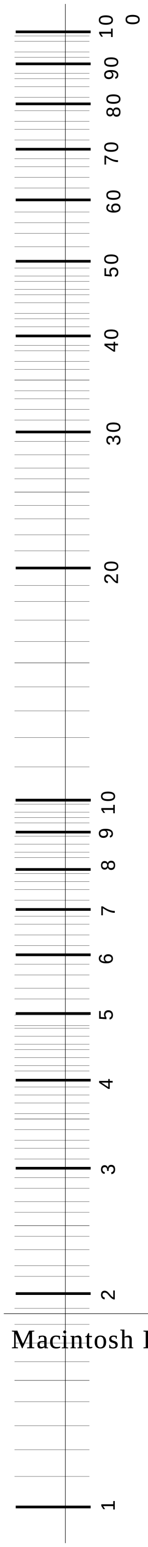
Find the "4" on the first scale.

Place the "1" on the second scale next to the "4" on the first scale

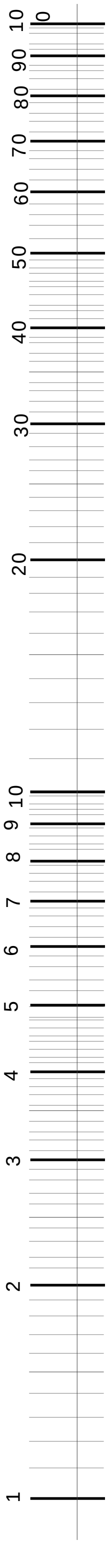
Find the "8" on the second scale,

And then read the answer on the first scale (finding "32" opposite the "4"

```
| 1------------------------------4--------------------------------------------------32       | 32---------------
|                                1----------------------------------------------------8      |
|                                                                                            |
| <Magnitude corresponding to 4>                                                             |
|                                  <Magnitude corresponding to 8>                            |
| < Magnitude_ _corresponding_ _to 32 >                                                      |
```



Two-Scale Slide Rule:
Tear into two strips and lay them edge to edge



The slide rule creates intervals for 1, 2, 3 that are the correct intervals when numbers are related by multiplication. Check it: Use a ruler to measure the distance between 1 and 4 on the slide rule and then measure the distance between 1 and 8 on the slide rule. You will find that the physical distance from 1 to 8 is 50% larger than the physical distance from 1 to 4 just as, in the exponents above, the exponent for 8 (as in $2^3 = 8$) is 50 percent larger than the exponent for 4 (as in $2^2 = 4$).

Linear Relations Using Logs

Interpreting *Log* y = mx+b

And now, finally, what happens if well-behaved form of a variable is its logarithm and you actually have to logarithms with data? You've got a variable: It's asymmetrical in its original form. But when you transform it, using logs, it is symmetrical. You look at the variable: It's dollars (an amount in Tukey's language) — you expect it to need logs. You plot the variable with respect to another variable and the result is definitely not linear — but when you plot its logarithm with respect to the other variable the result is definitely (close to) linear. So, the message is — "use logs", but how do you interpret such things?

To discuss the "meaning" of something whose logarithm appears to be a linear function of another variable — called a semi-log (or semi-logarithmic) relation — there is no trick: You have all of the mathematical tools, it's just a matter of using them.

So, here's what you've got, a semi-logarithmic linear relation

$$\ln Y = m \quad X \quad + \quad b$$

I know from simple math that if the semi-log relation is true, in nature, then an exponential relation is also true in nature. The second equation helps to interpret the first:

If $\ln y = mx + b$

then $y = e^{mx+b}$

Intercept

First, what is b? In the semi log equation, b is the intercept. Taking the algebra a step further, the form of the equation in dollars is

$$y = B\,e^{mx}, \text{ where } B = e^{b}$$

This demonstrates that the addition of b in the semilog relation implies a proportionality to anti-log b in the second equation. It says that y *is proportional* to an exponential function of x, where the proportionality, upper-case B, is the anti-log of lower-case b in the linear equation.

Slope

Now, second, what is m? In the linear form, of course, it means that *ln* y increase up by m units for each unit increase in x — that's what a slope says in a linear relation. If that is what happens in the logarithmic equation,, what happens in the second equation for plain dollars — without logs? There is no secret to answering the question. You just add 1 to x and see what happens.

So, in simple terms I ask again, what happens as x goes up by one unit?. I add one and look at the results

$$y = Be^{mx}$$
$$y' = Be^{m(x+1)}$$

That express the new value of y as y', corresponding to an x that is increased by 1. Simplifying the second equation, I get

$$y' = B\ e^{mx+m}$$

Breaking the exponential factor into two factors instead of just one, that is equivalent to

$$y' = B\ e^{mx}\ e^{m}$$

And now I recognize that the first part of the stuff on the right matches the original equation. So comparing y' to y the result depends on m. That tells me how to interpret m.

$$\frac{y'}{y} = \frac{Be^{m(x+1)}}{Be^{mx}}$$
$$\frac{y'}{y} = e^{m}$$

Ah, the effect of a unit increase in x is to multiply y by the value of e^m.

So suppose that m is a number like .03? I just pull out my calculator and figure out the value of the anti-log of .03. Or, if I am using logs base e, then I can remember the logs without computing them, remembering that exp(m) is equal to 1.03. Which means — here's the payoff, that y gets *multiplied* by 1.03 every time x *adds* 1 to its value.

And finally, we wipe out the traces of what we've done and the way we actually analyzed the data, by using percentages —

because people feel comfortable with percentages — even though they aren't much use when you're doing the real work — and I announce: "Y increases at 3% per annum.)

Scatter

Finally, because this is data analysis, there is the third property of a data analyst's line: The scatter.

The residuals are departures from the semi-log equation, log (y) = mx + b.

The trick is to put the residuals into a form that uses the words "plus or minus". If the residuals represent error, and nothing more, then the distribution of the residuals will be symmetrical around a mean of zero. So, for example, we can compute the standard deviation of the residuals and say, that the residuals have an average of zero "plus or minus" two standard deviations. Or, we can make the corresponding statement with medians and quartiles, saying that the residuals have an average of zero, "plus or minus" the number corresponding to the distance between the median and the quartiles (*either* quartile, since the distribution is symmetrical).

Then the corresponding statement in y (rather than log y) is straightforward (except that the standard deviation in units of log y is not the same as the standard deviation in units of y — so we avoid the word standard deviation. So: The residuals show deviation within a factor of ___, where you fill in the blank with the antilog of two standard deviations of y. Or, using quartiles: Fifty percent of the predicted values lie within a factor of ___ above the predicted values of y and a factor of ___ below the predicted values, where you fill in the first blank with the anti-log of the high quartile of the residuals and you fill in the second blank with the anti-log of the low quartile of the residuals. Or, using the inner fences: With few exceptions the predictions lie within a factor of __ above the predicted values and a factor of __ below them, where the blanks are filled in with anti-logs of the corresponding values of the fences of the residuals for log y.

U.S. Population: Not the Work, Not the Report, but the Thinking

I

My target for the day is the data describing the growth of the population of the United States. I want to derive a summary of the growth rate. I want to get an overview of the processes that generate it. Here are the data, Figure 1,

	Census Date	Resident Population					
	Conterminous U.S.	(Note 1)					
1790	Aug-02	3,929,214		1920	Jan-01	105,710,620	
1800	Aug-04	5,308,483		1930	Apr-01	122,755,046	
1810	Aug-06	7,239,881		1940	Apr-01	131,669,275	
1820	Aug-07	9,638,453		1950	Apr-01	150,697,361	
1830	Jun-01	12,866,020		1960	Apr-01	178,464,236	
1840	Jun-01	17,069,453					
1850	Jun-01	23,191,876		United	States		
1860	Jun-01	31,443,321		1950	Apr-01	151,325,798	
1870	Jun-01	39,818,449	Note 2	1960	Apr-01	179,823,175	
1880	Jun-01	50,155,783		1970	Apr-01	203,302,231	Note 3
1890	Jun-01	62,947,714		1980	Apr-01	226,545,805	
1900	Jun-01	75,994,575		1990	Apr-01	248,709,873	
1910	Apr-15	91,972,266					

Figure 1
United States Population: 1790 to 1990

Note 1: Excludes Alaska and Hawaii. Note 2: Revised to include adjustments for under numeration in southern states; unrevised number is 38558371. Note 3: Figures corrected after 1970 final reports were issued. From *Statistical Abstract of the United States*, 1992, No. 1. Original: U.S. Bureau of the Census, U.S. Census of Population: 1920 to 1990, vol. 1; and other reports.

I am thinking: Populations grow exponentially, which means that each year's growth is proportional to the preceding year's population. How do I know that? Truth is, I don't. But that is what all sorts of Malthusian folklore babbles about, so when I look at a population, I think growth *rate* (percentage) and think of the summary as the average growth rate. It takes people to make people, so at any time the growth of a population "should be" proportional to the size of the population. Do I believe that? No. I'm skeptical. If it were really obvious, if data behaved as data are "supposed to behave", there would be no need to analyze it. That's my thinking about "process" — a rudimentary hypothesis: growth in proportion to size.

That being said, I can make the first rough estimate in my head: The population doubled about six times in two hundred years. Doubling six times in two hundred years implies doubling one time in about 33 years, if the process was constant. And doubling in 33 years implies an annualized rate of increase of about 2% per annum. There is my first description, untested, of the growth process and growth rate for the United States. ("Rule of 70:" Divide 70 by the rate to estimate doubling time. Or, divide 70 by the number of years to get the rate. So, $70/33 \approx 2.1 \approx 2$: the population is doubling at about 2% per annum).

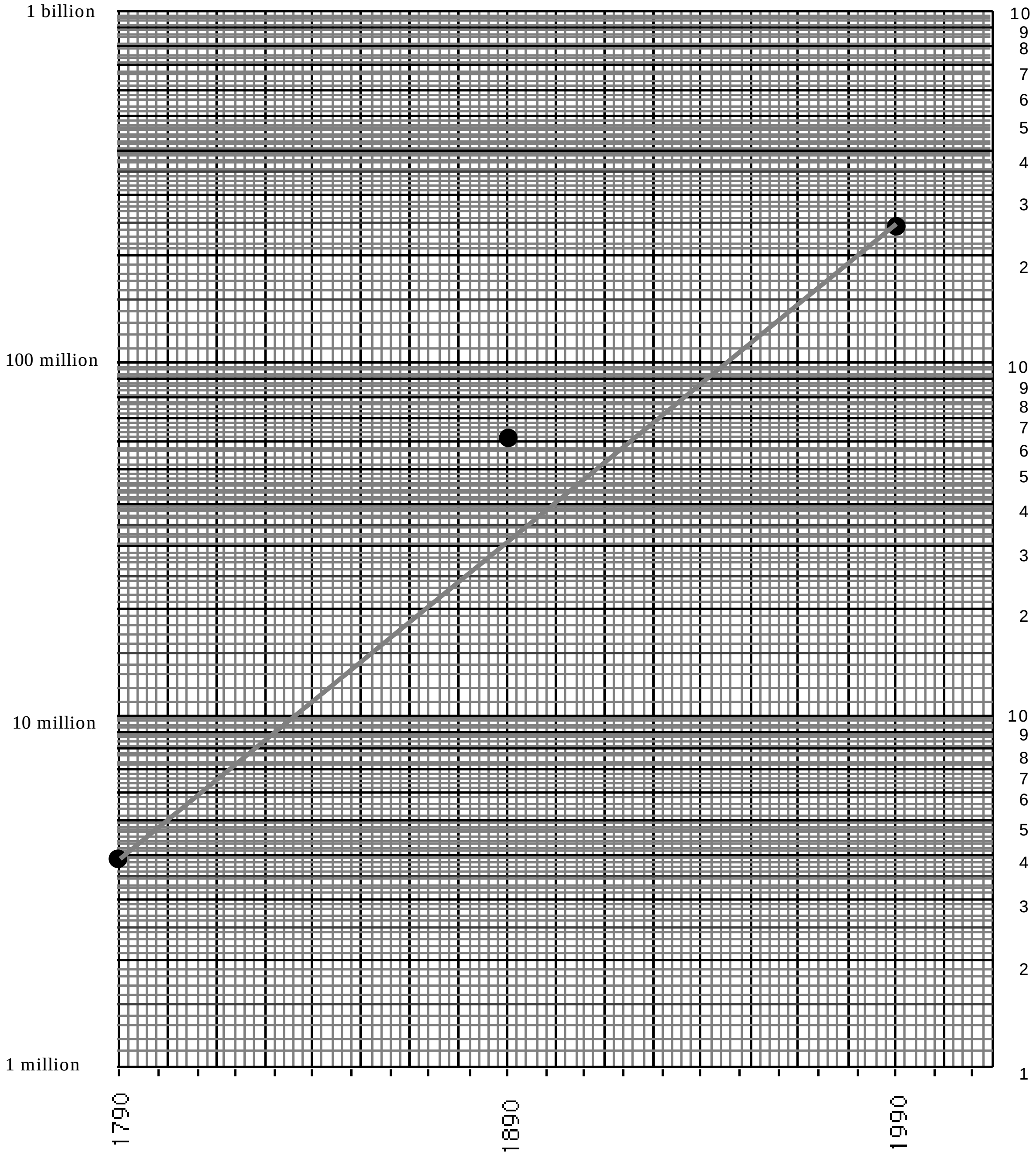
That's my first estimate. It gives me an order of magnitude to think about for all the rest: The annual rate of increase is about 2% per annum. Next, I'll graph it.

What kind of a graph? *Because* I'm thinking "grows by the same percentage each year", I want a kind of graph that is capable of falsifying this hypothesis if the hypothesis is false. I want a kind of graph that will look linear if that hypothesis is correct — but that will look non-linear if the hypothesis is false. If it is false, that will lead me back to re-examine the hypothesis "grows by the same percentage each year". Note: I'm not graphing because I like graphs, nor because that's "the next step". I'm graphing in order to see the actual rate (percentage growth) and I'm graphing so that if it is not behaving that way, then the graph will make it obvious.

So, again, what kind of a graph? Graphing it on ordinary graph paper (graphing the population counts) would be irrelevant — that kind of graph has nothing much to do with what I just said: On ordinary graph paper the growth of the U.S. population is surely going to be non-linear — whether or not my simple hypothesis is correct. So plotting the graph of U.S. population on ordinary graph would not advance my knowledge relative to my hypothesis — an ordinary graph would show that I was not thinking or not thinking clearly. — Waste of time.)

Ah, by contrast: Graphing it on semi-log graph paper, a straight line in *the semi-log graph* would show that my idea was consistent with the data — and failure to find a straight line would show inconsistency — I could learn something from that. So, using semi-log paper, and looking for a line will teach me something relative to my hypothesis.

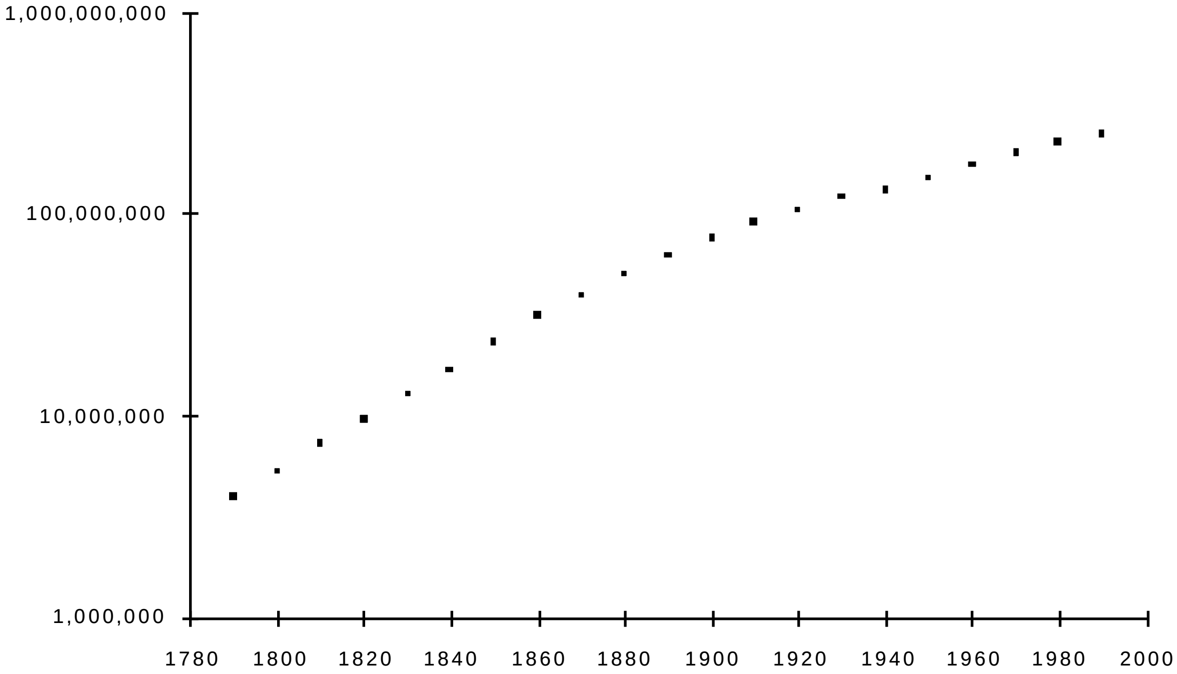
So, semi-log: And truth is, I can learn most of what I want to know by graphing exactly three points: One data point at the left. One data point at the right. And one data point in the middle. The data point in the middle provides the quickest way to test whether or not the data follow a straight line in their course between the first data point and the last. So


1 billion
100 million
10 million
1 million
10
9
8
7
6
5
4
3
2
10
9
8
7
6
5
4
3
2
10
9
8
7
6
5
4
3
2
1
1790
1890
1990

Oops: Those three points are not co-linear, not even close. The straight line tells me that my idea, constant proportional growth, would lead to a population of 30 million in 1890. But the data show 60 million in 1890— off by a factor of 2. So the idea is wrong — and I've learned something. Thomas Malthus can tell us that population grows exponentially while resources grow linearly, meaning that human populations will outrun the resources that feed us — with disastrous results. And lots of people can think about the great end-of-the-world implications that follow from Malthus' proposition. *Meanwhile* — we've checked the first part of his proposition against the U.S. data And? Its not true: I had an idea, a hypothesis. I framed it in a falsifiable way. And it was false.

One hypothesis — gone. Now, back to thinking. *Idea*: Could I be making too much out of a single point? No, implausible: The hypothetical value, assuming a constant rate of exponential growth missed the true value by a factor of 2, I can't rescue that idea by invoking "variability". *Idea*: Real human populations grow and shrink by immigration and emigration, as well as by biological reproduction. I might want to find separate data on these processes. The immigration idea sounds good, although it would require more data — easily obtained. *Idea*: Forsaking Malthus, I remember something about "demographic transition". That's what is supposed to separate the "first world" from the "third world": After industrialization, and a little taste of prosperity (like eating regularly and seeing health of children improve so that they can survive to become adults), people reduce their family size and increase the age at which they begin to bear children. After industrialization there is "supposed to be" a transition. Birth rates are "supposed to" drop — presumably because children get re-classified. Children change from being an asset (as a source of free labor to parents who live off the land) to being a drag on their parents' income (which is derived from employment away from the household).

That gives me a few ideas to work with, too many perhaps, and for now I'm going to just look at the graph: This time I am just "fishing" to see if it gives me ideas.

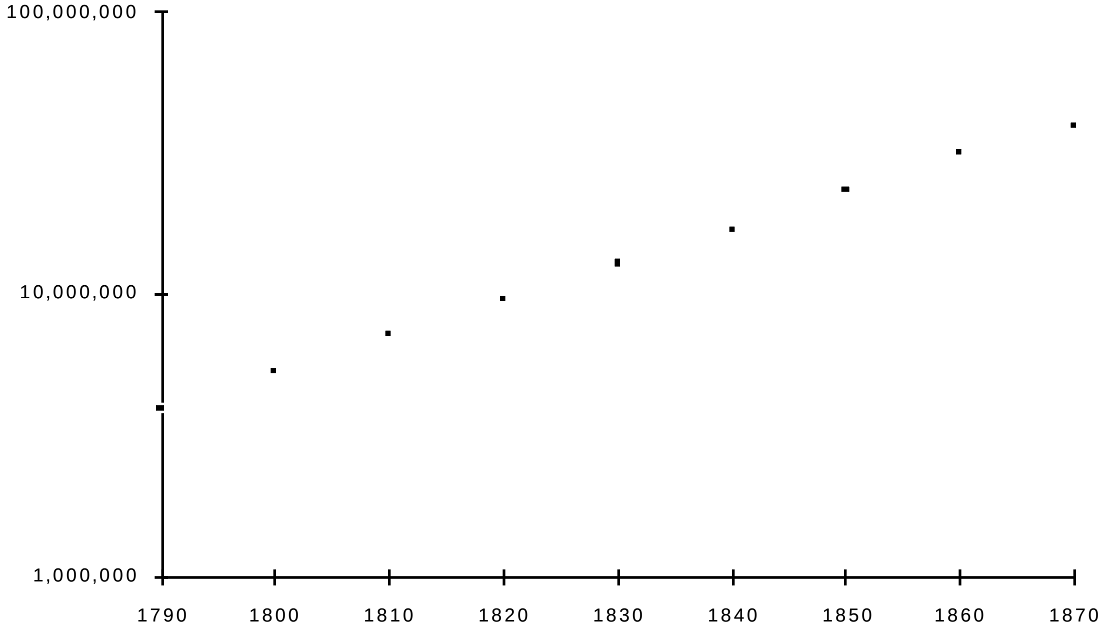


Well, it looks “sort of” linear through the first half. It bends. And then it looks “sort of” linear in the second half at a lower rate (with a dip in 1940). Was there a “demographic transition” between the first half and the second? Could be: U. S. industrialization is really supposed to have “taken off” in the 1870’s to 1890’s — transcontinental railroads, unified markets, telegraph, steel, oil, cartels, ... Let me start easy: Does the first half show a constant annual rate, disrupted in the late 19th century? That will not give me a falsifiable statement about the relatively sophisticated idea of a “demographic transition.” But I needn’t jump that far this quickly: I can put that off while I check the simpler statement “constant annual rate, disrupted in the late 19th century”.

So let me try 1790 through 1870. I want to graph it again. First I’ll do the semi-log graph. Why? to get a close-up” looking for non-linearity. If it is still plausible that the relation is linear, then I’ll look at the graph of logs to get an estimate of the slope and intercept. And then

I'll switch from the logs to the residuals. That's what I really want to see: the residuals, asking: "Do the residuals (for the early years) show a serious departure from linearity (on the semi-log graph)?"

So, I'll get the easy close-up traced on 2 cycle semi-log paper.



Looks good enough. Now, I'll invest the time to compute the logarithms numerically, computing the logarithms that were done automatically, or implicitly, by the semi-log graph paper. Why use the numbers? Because I want to see the residuals and to "see" them I have to compute them. And since I am going to put numbers on the logs, which logs? I'll use natural logs, logarithms base *e*, because, with base *e*, it is easy to recognize annual rates like 1 and 2 percent — which come out as .01 and .02 in logs base *e*..

So, setting-up my computation

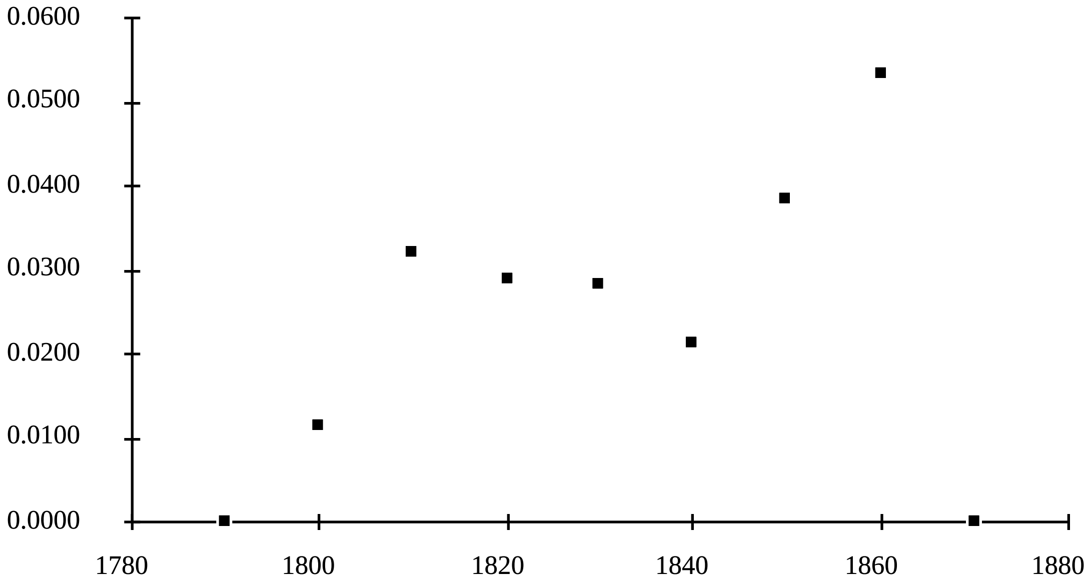
	slope	0.028948637		
	intercept	15.1839		
		Linear Prediction	Residual	Absolute Resid
1790	15.1839	15.1839	0.0000	0.0000
1800	15.4848	15.4734	0.0114	0.0114
1810	15.7951	15.7629	0.0322	0.0322
1820	16.0813	16.0524	0.0289	0.0289
1830	16.3701	16.3418	0.0283	0.0283
1840	16.6528	16.6313	0.0215	0.0215
1850	16.9593	16.9208	0.0385	0.0385
1860	17.2637	17.2103	0.0534	0.0534
1870	17.4998	17.4998	0.0000	0.0000
			Average	Average:

How to estimate slope and intercept? No problem: I just need a rough sketch. So, for a first estimate:

Slope: Last value minus first value, divided by 80.

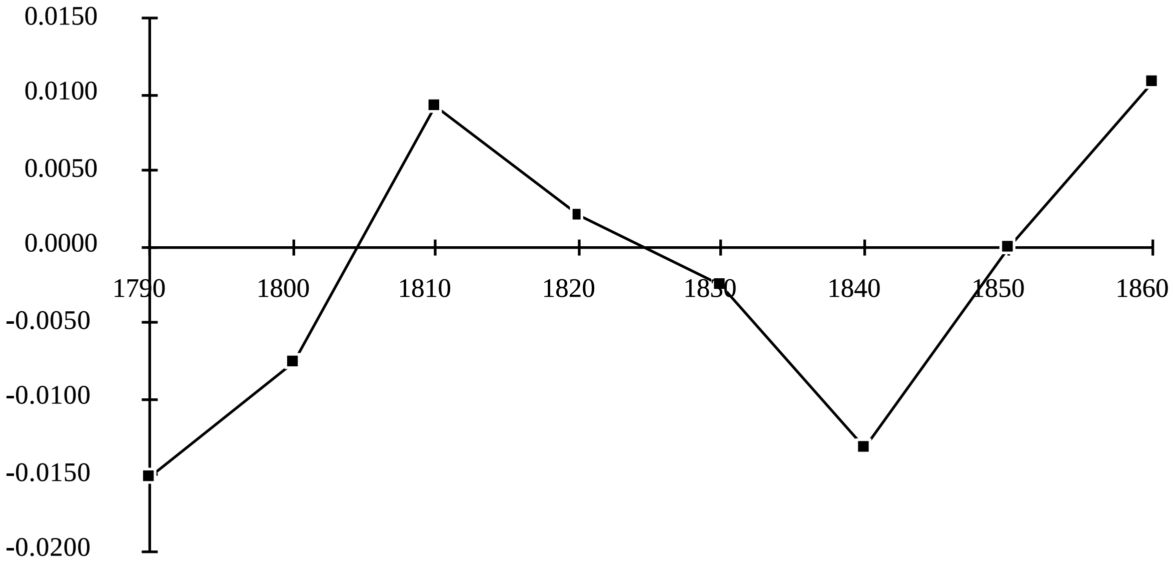
Intercept: I recalibrate "x" to be year since 1790. So "Year 0" corresponds to 1790. Then my estimate for the intercept is First value.

There's the graph. Strange: all the residuals are positive, and they are zero at both ends. That's a warning of curvature, but I'd better look more closely.



It surely isn't simple curvature. Maybe the last point, for 1870, is already out of the range for the early years and entering the range of industrialization. So let me drop the 1870 data point, and commit myself to the work involved in getting a more a serious attempt to get the slope and the intercept. I zero-in on it by looking at both the average residual and the average absolute residual: I fix the intercept so that the average residual is close to zero. Then I fix the slope to make the average *absolute* residual small. Then I fix the intercept so that the average residual is close to zero. Then I fix the slope to make the average *absolute* residual small. Then I fix the intercept so that the average residual is close to zero. ....

Year	Observed (In natural logarithms.)	Linear Prediction	Residual	Absolute Residual	Rank of Absolute Residuals
1790	15.1839	15.1990	-0.0151	0.0151	*8*
1800	15.4848	15.4924	-0.0076	0.0076	*4*
1810	15.7951	15.7858	0.0093	0.0093	*5*
1820	16.0813	16.0792	0.0021	0.0021	*2*
1830	16.3701	16.3726	-0.0025	0.0025	*3*
1840	16.6528	16.6660	-0.0132	0.0132	*7*
1850	16.9593	16.9594	-0.0001	0.0001	*1*
1860	17.2637	17.2528	0.0109	0.0109	*6*
1870	17.4998	17.5462	-0.0464	0.0464	*9*
	slope	0.02934			
	intercept	15.19900	Average Residual:	Average Absolute Residual:	Median Absolute Residual
			–0.002016972	0.007587886	.0093
	exptl of slope	1.02977	Exponentiated:	1.0076	1.0093
	exptl of inter	3,988,796			



Oops, that bothers me: That kind of cycling is exactly what you get when you are thinking about something wrong. (Although it can also be what you get from random numbers. I'm worried that while I have chosen to think about it in logs, my choice of this log form may not be valid. And why logs? Which is to say, why proportions? Do I really believe that populations grow in proportion to the number of people in the population? On second and third though, prompted by the facts, I'm not so sure about that. Proportional to the number of women is probably closer to the mark, but just barely: The point being that people are not yeast, one indistinguishable from another, all happily reproducing in the presence of nourishment and warmth. Human populations don't grow like that. A large part of the human population is not even "at risk" for reproduction, indeed the most rapidly increasing age cohorts of the population (the older cohorts), are little involved in reproduction. Got to think more carefully, about "proportional growth". Actually, now that I think about it, I'll bet that if we imagine that humans could live forever, our numbers would not increase out exponentially at all — the population growth is more likely to look linear, as the child bearing population size becomes stable in size (and a steadily decreasing part of the total population. So, “populations grow in proportion to their present size” is sloppy thinking. More accurate to say, we customarily *measure* the growth of populations by reporting the growth in proportion to previous size. Whether or not that growth rate is constant, or whether the growth is proportional to size — those are empirical questions.

For the moment, O.K., whatever that is, whether it is a cycling of some sort, or just an up and down — I have stripped the signal present in these data so that what’s left, the residuals, is small — perhaps too small to support my complicated theories that might have been built on them: How large are the residuals? The magnitudes of the residuals are, at worst, about 1.5 percent compared for these 10th year observations. With an annual growth rate of approximately 2.9% per year — that means my 10th year observations are off by a fraction of a single year’s growth. That seems small. So, summing it up: 1790 to 1860, the average annual growth rate is about 2.98 percent Applied to the seventy year period of these data, this growth rate predicts the population with a median error of less than 1% (using the median absolute residual, .0093, exponentiating it to 1.0093, converting it to a percent at 0.9%, and reporting it as approximately 1%). There is a suggestion of cycling, relative rapid growth, 1790 to 1820, relative slow, 1820 to 1870, then up. (Editing myself again: I’d Better get some number on those ups and downs. What numbers? People think in terms of annualized rates, so I want the annualized rates corresponding to these ten year periods. So, I compute the ten-year ratios, later to earlier, e.g., the ratio, 5,308,483 to 3929,214 for 1790 to 1800. That’s the multiplier for those 10 years. Then I estimate the annual ratio from the 10 year ratio by computing the 1/10 th power of the ten year ratio. And that’s the multiplier for the average single year, among the 10. That’s 1.0305.

1790	3,929,214	
1800	5,308,483	1.0305
1810	7,239,881	1.0315
1820	9,638,453	1.0290
1830	12,866,020	1.0293
1840	17,069,453	1.0287
1850	23,191,876	1.0311
1860	31,443,321	1.0309

So, 1790 to 1820, about 3.0%. 1820 to 1840, about 2.9% Those are the top and the bottom of one of the "cycles" I was seeing on the graph of the residuals That's small stuff (small difference) Maybe it is a pattern, it "looks" that way, but the difference between the rapid, 1790 to 1820, and the slow, 1820 to 1840, is too small for me to worry about, a difference of maybe one tenth of one percent, about 10,000 people on the 1820 population of 10 million.

That's my thinking. Without the thinking there is no reason for me to choose one graph or one set of numbers in preference to some other graphs and numbers. Without the thinking there is no reason for logs or not logs, for residuals, or something else.

Exercise:

Complete the Analysis:

Analyze the U.S. Population data for 1870-1890.

Write it up for 1790 - forward.

How Things Go Wrong

Data is equal to signal plus noise. That is the key. We use the key by examining the residuals, hoping that they will look like noise. If the residuals look like noise then, indirectly, they confirm that the hypothesis was a true representation of the signal.

In most examples even where a linear hypothesis is good, the first look at the residuals shows some straight forward evidence of a pattern. Until the intercept is right (in the hypothesis) the average residual is not zero. Until the slope is right (in the hypothesis) the residuals show a slope.

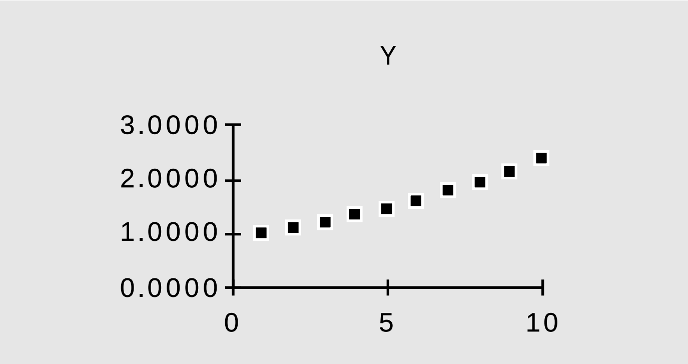
But suppose that the linear hypothesis is dead wrong. Then no patching of the intercept or the slope that will render the residuals patternless. If the linear hypothesis is wrong, what will the residuals look like? Logically the question would not seem to allow a general answer: There is no logical limit to the theoretical equations by which nature may choose to govern the relations among variables. And so it would seem that there is no end to the theoretical patterns which may show up in the residuals when the hypothesis is wrong.

But, in practice, things are not so bad. There may be no mathematical limit to the theoretical equation which nature may use, but, in practice, nature is rarely as perverse as is mathematically possible. I am going to demonstrate what can happen by committing a willful act of stupidity. Watch the residuals.

I'm going to ask you to work with me in a little exercise in curve fitting "by the book". All perfectly straightforward. Here are my data. You and I can see that these numbers are quite orderly. There is a clear system to the sequence 1, 1.1, 1.21, 1.331, etc. And they are a perfect candidate first for logarithmic transformation and then for a linear fit.

You know that but I'm going to forget it. And I'm going to forget the clever things I know about transformations. I'm just going to go at it as numbers, without complicating things by thinking too much. O.K. Here are the numbers, and here is their graph.

X	Y
1	1.0000
2	1.1000
3	1.2100
4	1.3310
5	1.4641
6	1.6105
7	1.7716
8	1.9487
9	2.1436
10	2.3579



I'm not sure what you see in this graph, or think you see in this graph, but let's suppose I come up with the observation that these numbers seem to be positive, and seem to exhibit a slope.

So here's the routine. I'm thinking of the schematic relation

Data = *Signal* + *Noise*

And practically I am matching it with the statement

Data - *Hypothesis* = *Residual.*

Then I'm going to subtract the hypothesis from the data, and "look" at the residuals — asking whether the residuals look like noise. If it does then my hypothesis, in the second equation, is a good approximation to the signal, in the first.

So I’ll start simple, very simple, with a hypothesis that says nothing at all. So, my “residuals are everything that was in the data, I’ve explained nothing.

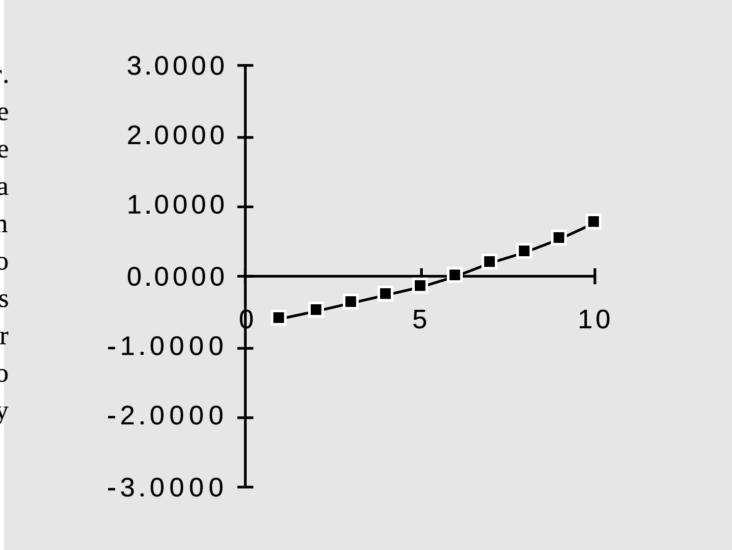
	Slope	0		
	Intercept	0		
X	Y	Y predicted	Residuals: observed Y -expected Y	Absolute Values of Residuals: |obs-exp|
1	1.0000	0	1.0000	1.0000
2	1.1000	0	1.1000	1.1000
3	1.2100	0	1.2100	1.2100
4	1.3310	0	1.3310	1.3310
5	1.4641	0	1.4641	1.4641
6	1.6105	0	1.6105	1.6105
7	1.7716	0	1.7716	1.7716
8	1.9487	0	1.9487	1.9487
9	2.1436	0	2.1436	2.1436
10	2.3579	0	2.3579	2.3579
		Average	1.5937	1.5937

What I see in these residuals is that they are all positive. So I want to transfer the information “this is positive” out of the residuals and into the hypothesis. It looks positive by about 1.59, (the average value of the “Other”) . I’ll add this to my hypothesis:

X	Y	Y predicted	Other: observed Y-expected Y	Absolute Value: |obs-exp|
1	1.0000	0	-0.5900	0.5900
2	1.1000	0	-0.4900	0.4900
3	1.2100	0	-0.3800	0.3800

4	1.3310	0	-0.2590	0.2590
5	1.4641	0	-0.1259	0.1259
6	1.6105	0	0.0205	0.0205
7	1.7716	0	0.1816	0.1816
8	1.9487	0	0.3587	0.3587
9	2.1436	0	0.5536	0.5536
10	2.3579	0	0.7679	0.7679
		Average	0.0037	0.3727
	Slope	0		
	Intercept	1.59		

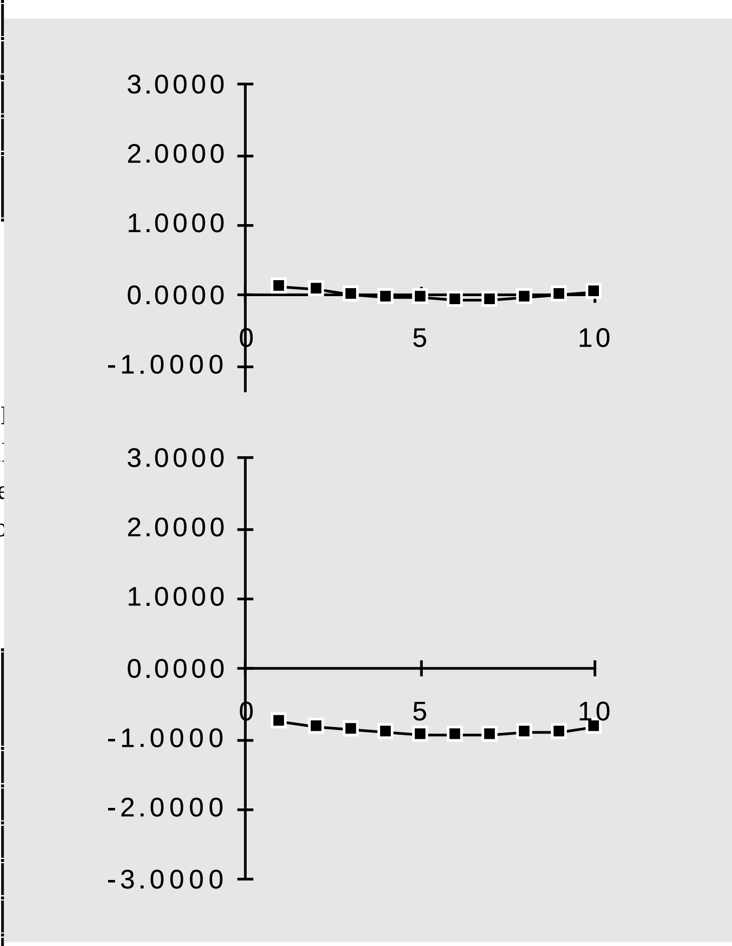
That's progress: The residuals are smaller. Checking both the average residual and the mean absolute size of the residuals, the residuals are smaller. But, visibly, there is a positive slope to these residuals: They go from about -.6 to about +.8 as x goes from 1 to 10. So a good estimate of the slope in the residuals would be approximately about (.8 + .6)/9 or approximately 1.4/9, or approximately .16. So I'll add this slope to the hypothesis (thereby subtracting the slope from the residuals).



Moving the slope into my hypothesis:

X	Y	Y predicted	Other: observed Y-expected Y	Absolute Value: |obs-exp|
1	1.0000	1.75	-0.7500	0.7500
2	1.1000	1.91	-0.8100	0.8100
3	1.2100	2.07	-0.8600	0.8600
4	1.3310	2.23	-0.8990	0.8990
5	1.4641	2.39	-0.9259	0.9259
6	1.6105	2.55	-0.9395	0.9395
7	1.7716	2.71	-0.9384	0.9384

8	1.9487	2.87	-0.9213	0.9213
9	2.1436	3.03	-0.8864	0.8864
10	2.3579	3.19	-0.8321	0.8321
		Average	-0.8763	0.8763
	Slope	0.16		
	Intercep t	1.59		

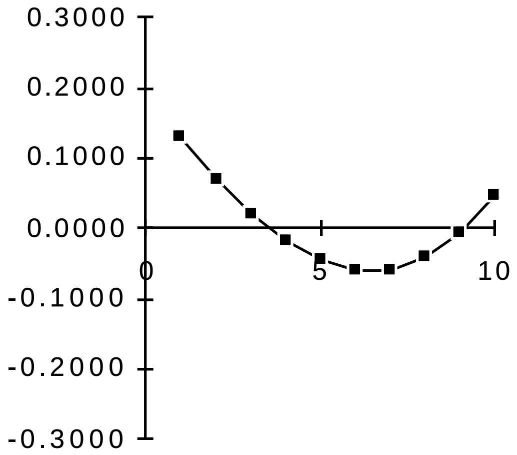


Ah, I got rid of the slope. But now I can see that my hypothesis about the signa is a little to large, too high, leaving negative residuals, about -.88. So I will move this too from the residuals to the hypothesis.

X	Y	Y predicted	Residuals: observed Y-expected Y	Absolute Value: |obs-exp|
1	1.0000	0.87	0.1300	0.1300
2	1.1000	1.03	0.0700	0.0700
3	1.2100	1.19	0.0200	0.0200
4	1.3310	1.35	-0.0190	0.0190
5	1.4641	1.51	-0.0459	0.0459
6	1.6105	1.67	-0.0595	0.0595
7	1.7716	1.83	-0.0584	0.0584
8	1.9487	1.99	-0.0413	0.0413
9	2.1436	2.15	-0.0064	0.0064
10	2.3579	2.31	0.0479	0.0479
		Average	0.0037	0.0498
	Slope	0.16		
	Intercept	0.71		

There: tiny residuals. The average deviation, using absolute values to check variation in either direction, is about .05. That’s small compared to the original 1.59, one thirtieth of the original residuals and small compared to the values I am trying to predict (which range from 1 to 2.36). So, is my work complete? Well, not quite. I can't really "see" what's left in this graph precisely because the stuff that's left is so small compared to the original scale. So, just to get a good luck at the residuals, let me change the scale of the graph and look again — just to be sure. And ...

Same numbers, expanded scale on their graph



That’s trouble. The residuals show a clear pattern. And thus, unraveling my verbal equations, the residuals in equation 2 do not look like “noise” as specified in equation 1. And that implies that my hypothesis in equation 2 does not look like the unknown “signal” in equation 1.

Pausing for the moment to emphasize the moral to this story, the pattern in this graph is lesson #1: When the hypothesis is wrong, dead wrong, the residuals often fall into one of two patterns, a simple curve, concave up or concave down.

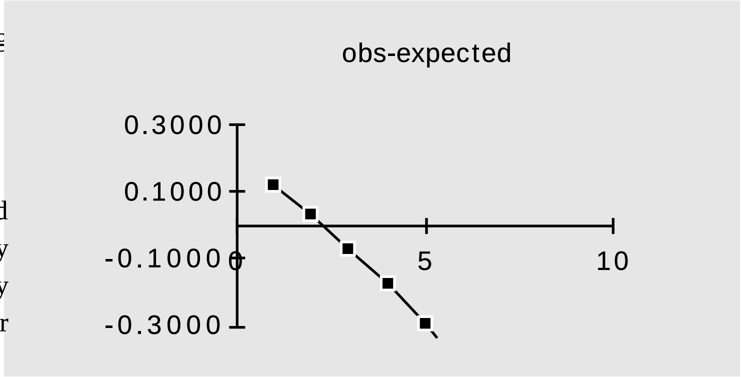
The errors may be tiny, but no matter. It is quite clear that the linear hypothesis does not describe the process. If this is the pattern of the residuals, then the linear hypothesis is wrong, numerically accurate, but dead wrong.

That’s lesson #1. Now, back to the analysis. What am I to do with this curving residual? What I should do is step back and think. But instead of flexing my intellectual muscle I am going to flex my arithmetical muscle and use the power of my (decidedly unintellectual) computer. Thinking about the arithmetic, the “obvious” answer is: Add something curvy to the hypothesis, matching or attempting to match the curviness in the residuals. All right, suppose I add a “quadratic” term, an x-squared term in addition to the existing linear term and the constant. If the linear equation

$$y = mx + b$$

did not work, then I will up the ante by adding another term. Switching notation, I will try

$$y = a_0 + a_1x + a_2x^2$$

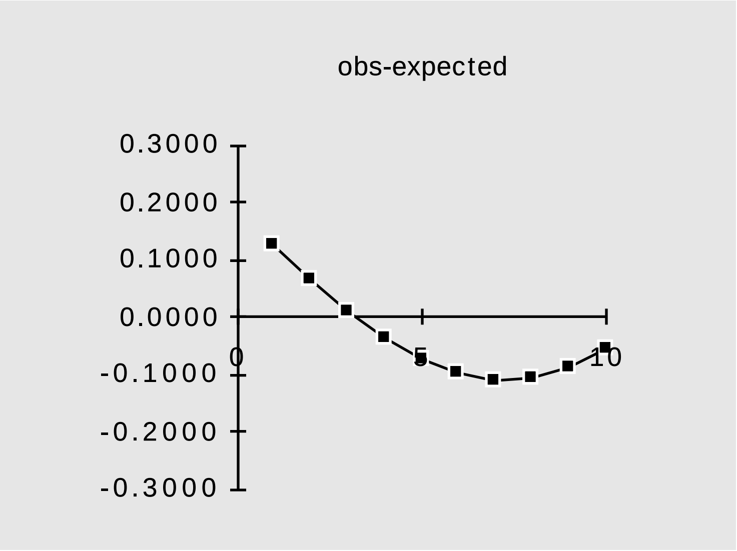


That is the new and more sophisticated equation (falsely sophisticated). I test it by looking at the residuals, the same way I test any hypothesis. Starting with a small estimate for the quadratic term, using $.01x^2$.

X	Y	Y predicted	Other: observed Y-expected Y	Absolute Value: |obs-exp|
1	1.0000	0.88	0.1200	0.1200
2	1.1000	1.07	0.0300	0.0300
3	1.2100	1.28	-0.0700	0.0700
4	1.3310	1.51	-0.1790	0.1790
5	1.4641	1.76	-0.2959	0.2959
6	1.6105	2.03	-0.4195	0.4195
7	1.7716	2.32	-0.5484	0.5484
8	1.9487	2.63	-0.6813	0.6813
9	2.1436	2.96	-0.8164	0.8164
10	2.3579	3.31	-0.9521	0.9521
		Average	-0.3813	0.4113
	Slope	0.16		
	Intercept	0.71		
	Quad term	0.01		

Small as it was it has messed up the residuals, not simplified them. So, I'll try something smaller:

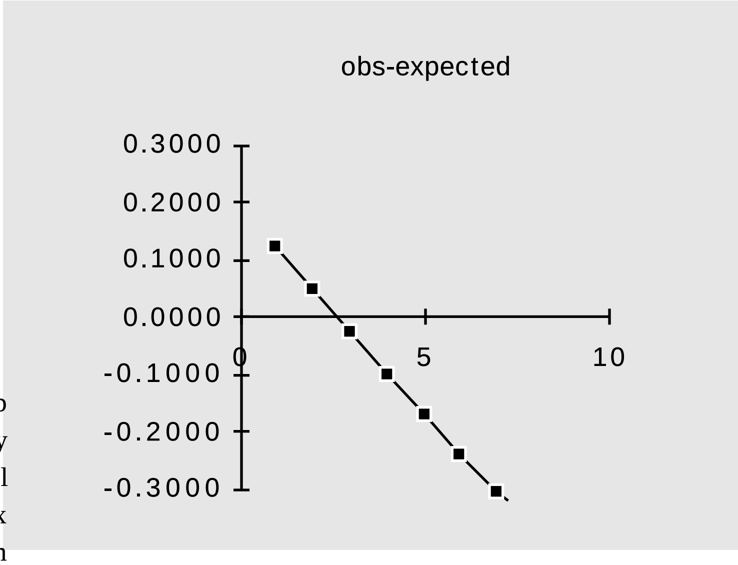
X	Y	Y predicted	Other: observed Y-expected Y	Absolute Value: |obs-exp|
1	1.0000	0.871	0.1290	0.1290
2	1.1000	1.034	0.0660	0.0660
3	1.2100	1.199	0.0110	0.0110
4	1.3310	1.366	-0.0350	0.0350
5	1.4641	1.535	-0.0709	0.0709
6	1.6105	1.706	-0.0955	0.0955
7	1.7716	1.879	-0.1074	0.1074
8	1.9487	2.054	-0.1053	0.1053
9	2.1436	2.231	-0.0874	0.0874
10	2.3579	2.41	-0.0521	0.0521
		Average	-0.0348	0.0760
	Slope	0.16		
	Intercept	0.71		
	Quad term	0.001		



Better. Right direction. But there is still a curve in the residuals. So let me put more curve into my hypothesis, subtracting it from these residuals.

X	Y	Y predicted	Other: observed Y-expected Y	Absolute Value: |obs-exp|
1	1.0000	0.875	0.1250	0.1250
2	1.1000	1.05	0.0500	0.0500
3	1.2100	1.235	-0.0250	0.0250
4	1.3310	1.43	-0.0990	0.0990
5	1.4641	1.635	-0.1709	0.1709
6	1.6105	1.85	-0.2395	0.2395
7	1.7716	2.075	-0.3034	0.3034

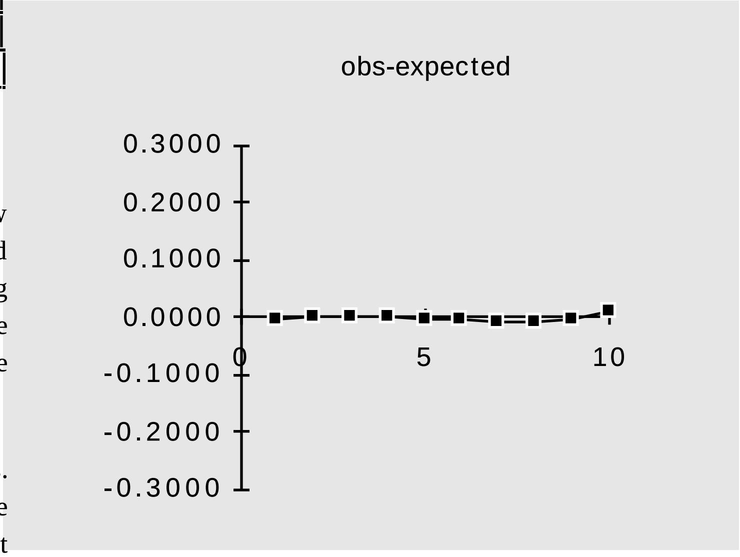
8	1.9487	2.31	-0.3613	0.3613
9	2.1436	2.555	-0.4114	0.4114
10	2.3579	2.81	-0.4521	0.4521
		Average	-0.1888	0.2238
	Slope	0.16		
	Intercept	0.71		
	Quad term	0.005		



That's looking straight. So I have to fix up the slope: I've got a negative slope to my residuals, so I have to make the hypothetical slope more negative. And then I will have to fix up the intercept, and then the quadratic term again, and then the slope again, and then the intercept again, and .... I get:

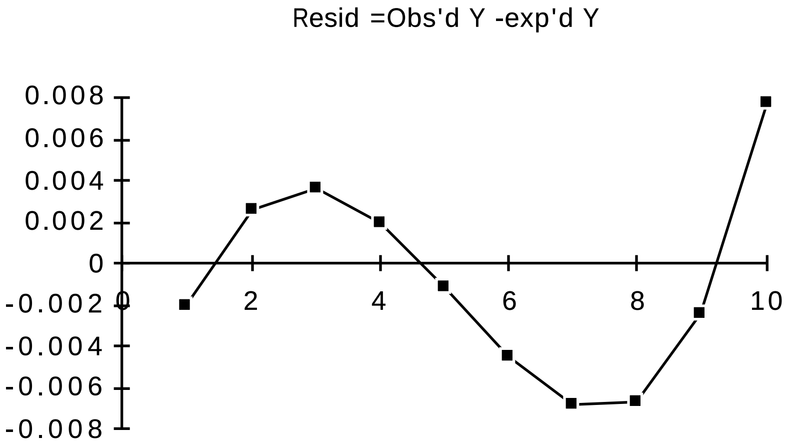
X	Y	Y predicted	Other: observed Y-expected Y	Absolute Value: |obs-exp|
1	1.0000	1.002	-0.0020	0.0020
2	1.1000	1.0973	0.0027	0.0027
3	1.2100	1.2062	0.0038	0.0038
4	1.3310	1.3287	0.0023	0.0023
5	1.4641	1.4647	-0.0006	0.0006
6	1.6105	1.6143	-0.0038	0.0038
7	1.7716	1.7774	-0.0059	0.0059
8	1.9487	1.9541	-0.0054	0.0054
9	2.1436	2.1444	-0.0008	0.0008
10	2.3579	2.3482	0.0097	0.0097
		Average	0.0000	0.0037
	Slope	0.075		

	Intercept	0.9202		
	Quad term	0.0068		



Now, look at that! Depending on how broadly those lines come out on the printed page, I've reduced my error to something that practically disappears within the breadth of the line I've used for the horizontal axis. Now, I've got it. Right?

Well, … let's look at the residuals. They are small, an order of magnitude smaller than they were in my best effort using the straight line (without the quadratic term). But then let me also improve the scale of my graph.



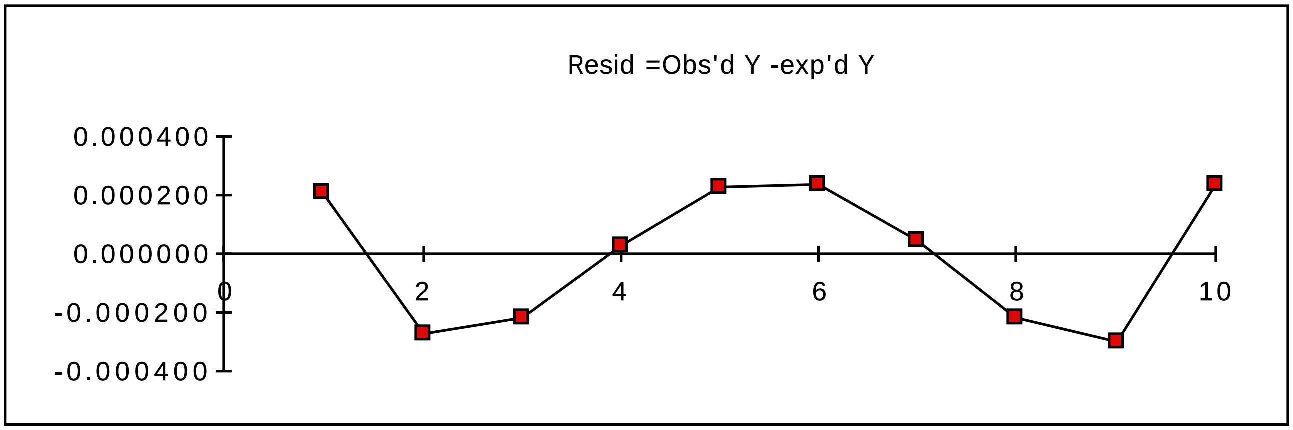
The residuals are tiny, plus or minus .005, give or take. No matter. The residuals show a pattern. The hypothesis is wrong (no matter how small the errors).

And suppose I'm a slow learner, perfectly capable of fitting a (supposedly) more sophisticated model

$$y = a_0 + a_1x + a_2x^2 + a_3x^3$$

Then nature will oblige by punishing my error with an even more interesting pattern among the residuals.

		slope	0.08875	Intercept	0.90742	Quadratic:	0.00338765
X	Y	Expected Y(X)	Resid =Obs'd Y -exp'd Y	Squared Residual		Cubic	0.00022400
1	1.000000	0.999784	0.000216	4.6592E-08			
2	1.100000	1.100268	-0.000268	7.1601E-08			
3	1.210000	1.210214	-0.000214	4.5923E-08			
4	1.331000	1.330968	0.000032	1.0067E-09			
5	1.464100	1.463874	0.000227	5.1302E-08			
6	1.610510	1.610274	0.000236	5.5711E-08			
7	1.771561	1.771514	0.000047	2.2407E-09			
8	1.948717	1.948937	-0.000219	4.817E-08			
9	2.143589	2.143887	-0.000298	8.8734E-08			
10	2.357948	2.357708	0.000240	5.7452E-08			
		mean	0.00	0.00000			
				0.00021650			



Some people will tell you that the sine qua non of science is prediction. But that is too simple a dictum to follow blindly. Here, with my cubic equation I have used the equation to predict or match the values of y, matching these data with precision that is so good that the remaining errors are beginning to get lost in the normal rounding errors made by my computer. I started with a “y” that ranged from 1 to 2.38. I’ve matched those numbers subject to errors which are less than 0.003 in absolute value. I’ve fit the data without attempting to understood it — which is a waste of time. I would rank this data analysis as overly mathematical, unnecessarily precise, technically difficult, totally lacking in insight, and dead wrong.

The Meaning of the Pattern

I have demonstrated this nonsense in such detail because this particular sequence of patterns among residuals will haunt you as you proceed in data analysis. This is what you get when there is an answer — but you are not approaching it correctly. In this case there is an answer: This is an exponential growth curve. It is linear in log y. But I approached it incorrectly when I chose to stay with y (without considering a transformation) and when I attempted to force a polynomial to fit the data.

What you are seeing at work in these residuals is a part of mathematics known to every student of the calculus: You are seeing certain aspects of "power series" at work. Power series and related methods are able to fit a polynomial equation to any sequence of numbers (preferable finite) to any standard of precision — provided you can accept a polynomial of sufficiently high degree.

For example, here is one power series for ln(1+x) in the range of this problem:

$$\ln(1+x) = x - \tfrac{1}{2}x^2 + \tfrac{1}{3}x^3 - \tfrac{1}{4}x^4 + \ldots (-1 < x < 1)$$

The point is simply that when you get it wrong, "thinking" the expression on the right instead of the simple expression on the left then what can happen is that when you fit a line, what remains may be dominated by missing terms in the 2nd power, 3rd power, and more. When you fit a quadratic, what remains may be dominated by missing terms in the 3rd power, 4th power, and more. And so on, until you run out of data. So when you see systematic variation among the residuals, stop, look to your hypothesis, not your computer.

Exercise

Lest you think that residuals of the sort I've described are a mathematical possibility but not a realistic concern, with data:

Plasticity of Wool -- from Tukey ***

U.S. Population:
A Second Analysis

Again, my target for the day is the data describing the growth of the population of the United States. I want to derive a summary of the growth rate. I want to get an overview of the processes that generate it. This time I am going to construct a distinctly different analysis, as compared to the first.

How is it that I can construct two seriously different analyses of one set of data? I can do that because there is something detached about a text or a course on data analysis — it is detached from the actual research. Limited to these numbers, without recourse (for the moment) to other data (data on birth rates, death rates, life expectancy, age distribution, occupational statistics, immigration rates, emigration rates, ...) I can follow divergent thoughts from one set of data without the commitment and the resources by which serious research would formulate hypotheses and choose among them. Once again, here are the data, Figure 1,

	Census Date	Resident Population	
	Conterminous U.S.	(Note 1)	
1790	Aug-02	3,929,214	
1800	Aug-04	5,308,483	
1810	Aug-06	7,239,881	
1820	Aug-07	9,638,453	
1830	Jun-01	12,866,020	
1840	Jun-01	17,069,453	
1850	Jun-01	23,191,876	
1860	Jun-01	31,443,321	
1870	Jun-01	39,818,449	Note 2
1880	Jun-01	50,155,783	
1890	Jun-01	62,947,714	
1900	Jun-01	75,994,575	
1910	Apr-15	91,972,266	
1920	Jan-01	105,710,620	
1930	Apr-01	122,755,046	
1940	Apr-01	131,669,275	
1950	Apr-01	150,697,361	
1960	Apr-01	178,464,236	
United	States		
1950	Apr-01	151,325,798	
1960	Apr-01	179,823,175	
1970	Apr-01	203,302,231	Note 3
1980	Apr-01	226,545,805	
1990	Apr-01	248,709,873	

Figure 1
United States Population: 1790 to 1990

Note 1: Excludes Alaska and Hawaii. Note 2: Revised to include adjustments for under numeration in southern states; unrevised number is 38558371. Note 3: Figures corrected after 1970 final reports were issued. From *Statistical Abstract of the United States*, 1992, No. 1. Original: U.S. Bureau of the Census, U.S. Census of Population: 1920 to 1990, vol. 1; and other reports.

This time I am thinking simply that populations grow exponentially. Maybe, so let me test that. To test that I follow the usual procedure: I set up a graph such that if the hypothesis is correct then the graph will be linear and the residuals will be noise. For exponential growth that means computing the logarithms, fitting a line to the logarithms of population, and looking at the residuals. Because I expect growth rates on the order of 2 to 3 percent and residuals on the same scale, I will use logarithms base e. (For 1950 and 1960, the two estimates, with and without Alaska and Hawaii, differ by about half of one percent, a small but non-trivial difference. For the moment, I will use their mean.)

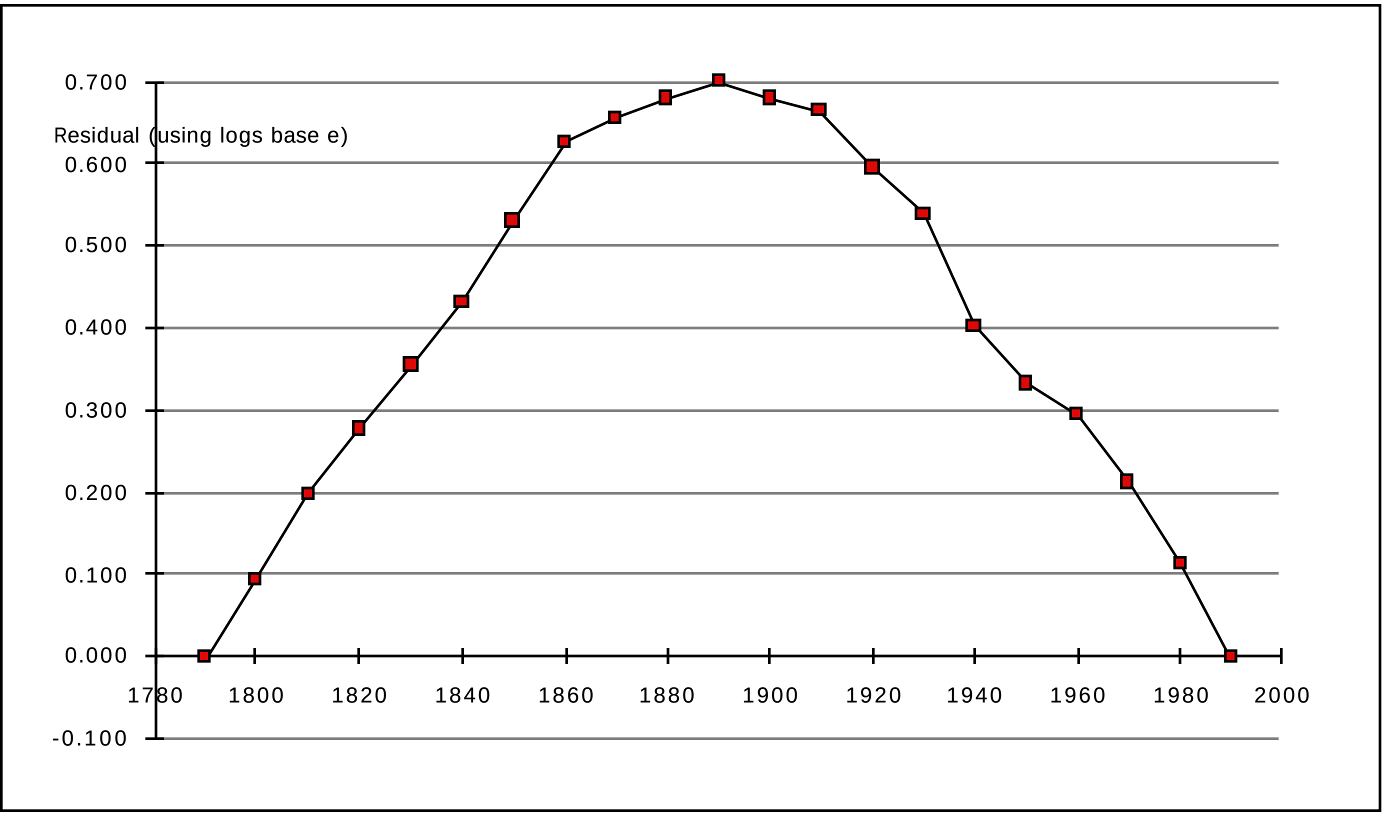
For calculation purposes I will count 1790 as year "0". That allows me to use the log of the population in 1790 ad a first estimate of the intercept. Then for an estimate of the slope I compute a "rise" from the difference between the logarithm of the 1990 population and the logarithm of the 1790 and I compute a "run" of 200 years, estimating a slope of

Intercept	15.184
slope	0.02073924

Year	Years after 1790		ln(Pop)	Expected	Residual	Sqd Residual
1790	0	3,929,214	15.184	15.184	0.000	0.000
1800	10	5,308,483	15.485	15.391	0.093	0.009
1810	20	7,239,881	15.795	15.599	0.196	0.039
1820	30	9,638,453	16.081	15.806	0.275	0.076
1830	40	12,866,020	16.370	16.014	0.357	0.127
1840	50	17,069,453	16.653	16.221	0.432	0.186
1850	60	23,191,876	16.959	16.428	0.531	0.282
1860	70	31,443,321	17.264	16.636	0.628	0.394
1870	80	39,818,449	17.500	16.843	0.657	0.431
1880	90	50,155,783	17.731	17.051	0.680	0.463
1890	100	62,947,714	17.958	17.258	0.700	0.490

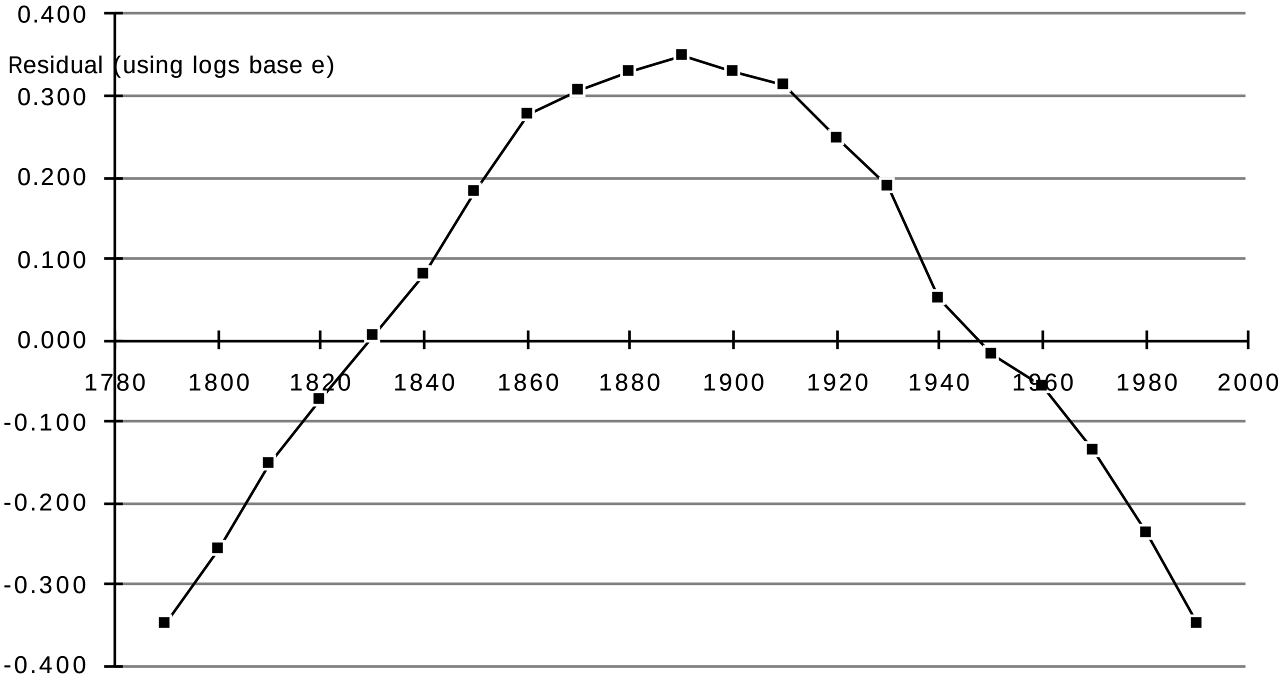
1900	110	75,994,575	18.146	17.465	0.681	0.464
1910	120	91,972,266	18.337	17.673	0.664	0.441
1920	130	105,710,620	18.476	17.880	0.596	0.355
1930	140	122,755,046	18.626	18.087	0.538	0.290
1940	150	131,669,275	18.696	18.295	0.401	0.161
1950	160	151,011,580	18.833	18.502	0.331	0.109
1960	170	179,143,706	19.004	18.710	0.294	0.086
1970	180	203,302,231	19.130	18.917	0.213	0.045
1980	190	226,545,805	19.238	19.124	0.114	0.013
1990	200	248,709,873	19.332	19.332	0.000	0.000

Mean Sqd Resid 0.21243886



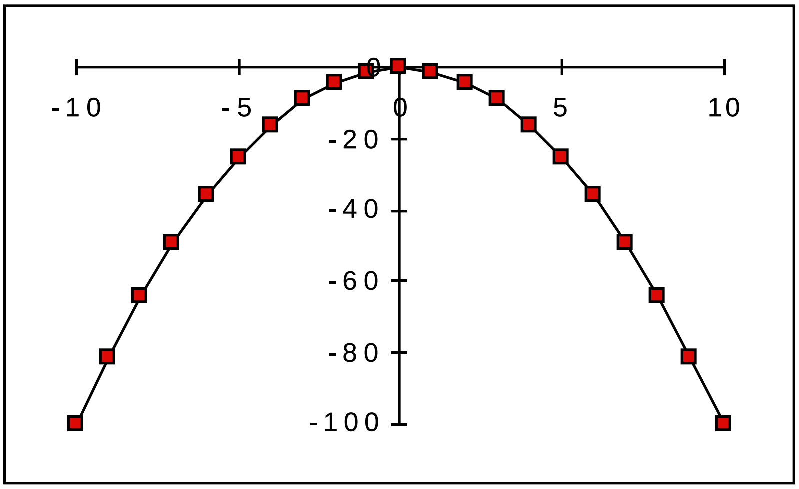
There is an obvious problem with this first intercept. So revising the hypothesis by adding 0.35 to the intercept:

			Intercept	15.534		
			slope	0.02073924 0		
Year	Years after 1790		ln(Pop)	Expected	Residual	Sqd Residual
1790	0	3,929,214	15.184	15.534	-0.350	0.123
1800	10	5,308,483	15.485	15.741	-0.257	0.066
1810	20	7,239,881	15.795	15.949	-0.154	0.024
1820	30	9,638,453	16.081	16.156	-0.075	0.006
1830	40	12,866,020	16.370	16.364	0.007	0.000
1840	50	17,069,453	16.653	16.571	0.082	0.007
1850	60	23,191,876	16.959	16.778	0.181	0.033
1860	70	31,443,321	17.264	16.986	0.278	0.077
1870	80	39,818,449	17.500	17.193	0.307	0.094
1880	90	50,155,783	17.731	17.401	0.330	0.109
1890	100	62,947,714	17.958	17.608	0.350	0.122
1900	110	75,994,575	18.146	17.815	0.331	0.109
1910	120	91,972,266	18.337	18.023	0.314	0.099
1920	130	105,710,620	18.476	18.230	0.246	0.061
1930	140	122,755,046	18.626	18.437	0.188	0.035
1940	150	131,669,275	18.696	18.645	0.051	0.003
1950	160	151,011,580	18.833	18.852	-0.019	0.000
1960	170	179,143,706	19.004	19.060	-0.056	0.003
1970	180	203,302,231	19.130	19.267	-0.137	0.019
1980	190	226,545,805	19.238	19.474	-0.236	0.056
1990	200	248,709,873	19.332	19.682	-0.350	0.123
					Mean Sqd Resid	0.05557615



Clearly, these residuals show a pattern and, therefore, the hypothesis is wrong. The fit of this line to the logs yields residuals ranging from -.35 to +.35, meaning that the ratio between the true populations and the populations that would be expected (were the hypothesis correct) range as high as the exponential of .35, which is 1.41, errors of 41% at the extremes.

But, look at the graph. It looks "sort of" quadratic. It looks like an almost straight line, rising. It bends. And then it looks almost straight, falling away. My mathematical repertoire tells me that quadratic equations can look like that. For example, here is a graph of the function $y = -x^{2,}$, graphed between x = -10 and x=10.



Is that mathematical pattern the pattern I've seen in these residuals? Let me hypothesize that it is, and then test it.

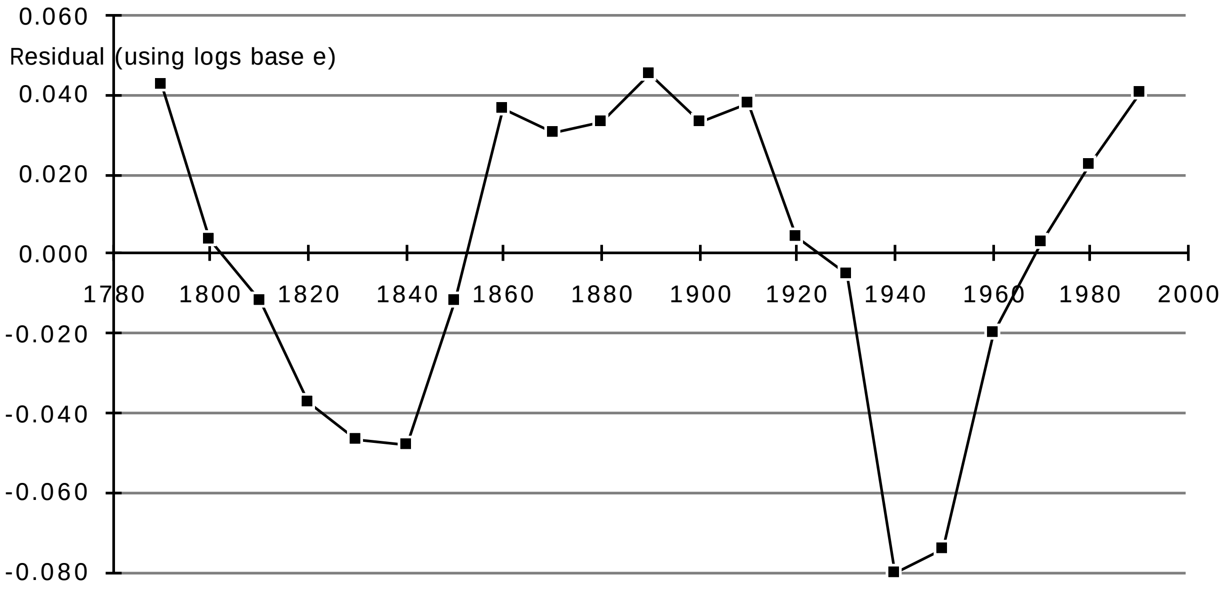
But first, never mind the mathematics, what would it mean if the hypothesis were correct? How would I interpret a quadratic equation and what would it tell me? My reflex is to search my memory for something analogous. And what I come up with is something from simple physics: The equation for the position of a particle moving in a straight line. If the particle begins at x_0 with velocity v (at time 0), and accelerates with acceleration a, then the descriptive equation is

$$x(t) = x_0 + vt + \tfrac{1}{2}at^2$$

That feeds my intuition: If a quadratic equation fits these data, thenthe coefficient of the quadratic term would be desribing acceleration or deceleration in the growth rate. That's good. I know already that the growth rate in 1790 is about one percent larger than the growth rate in 1990. The quadratic equation would express a hypothesis that the growth declined smoothly over the two hundred perios (not suddenly, circa 1890). So, I will be able to interpret it, if I need to.

So, back to my spread sheet, I keep adjusting my estimates of the coefficients, reducing the mean squared residual and I get:

			Intercept	15.1411		
			slope	0.0346704		
			Quadratic	-6.96E-05		
Year	Years after 1790		ln(Pop)	Expected	Residual	Sqd Residual
1790	0	3,929,214	15.184	15.141	0.043	0.002
1800	10	5,308,483	15.485	15.481	0.004	0.000
1810	20	7,239,881	15.795	15.807	-0.012	0.000
1820	30	9,638,453	16.081	16.119	-0.037	0.001
1830	40	12,866,020	16.370	16.417	-0.046	0.002
1840	50	17,069,453	16.653	16.701	-0.048	0.002
1850	60	23,191,876	16.959	16.971	-0.011	0.000
1860	70	31,443,321	17.264	17.227	0.037	0.001
1870	80	39,818,449	17.500	17.469	0.031	0.001
1880	90	50,155,783	17.731	17.698	0.033	0.001
1890	100	62,947,714	17.958	17.912	0.046	0.002
1900	110	75,994,575	18.146	18.113	0.033	0.001
1910	120	91,972,266	18.337	18.299	0.038	0.001
1920	130	105,710,620	18.476	18.472	0.004	0.000
1930	140	122,755,046	18.626	18.631	-0.005	0.000
1940	150	131,669,275	18.696	18.776	-0.080	0.006
1950	160	151,011,580	18.833	18.907	-0.074	0.005
1960	170	179,143,706	19.004	19.024	-0.020	0.000
1970	180	203,302,231	19.130	19.127	0.003	0.000
1980	190	226,545,805	19.238	19.216	0.022	0.001
1990	200	248,709,873	19.332	19.291	0.040	0.002
					Mean Sqd Resid	0.00144635



That is interesting. My first inspection sees two things: The residuals range from -.08 to +.04, which means that the population estimates are off by a maximum of 4 to 8 percent. Second, the residuals are flat circa 1890 — which is the point at which the earlier analysis hypothesized a break in the pattern.

In more detail, I see that the errors are, many of them, on the order of 4%, slightly larger but not dramatically larger thanthe errors visible in the earlier analysis. So, this description is "competitive" with the earlier analysis.

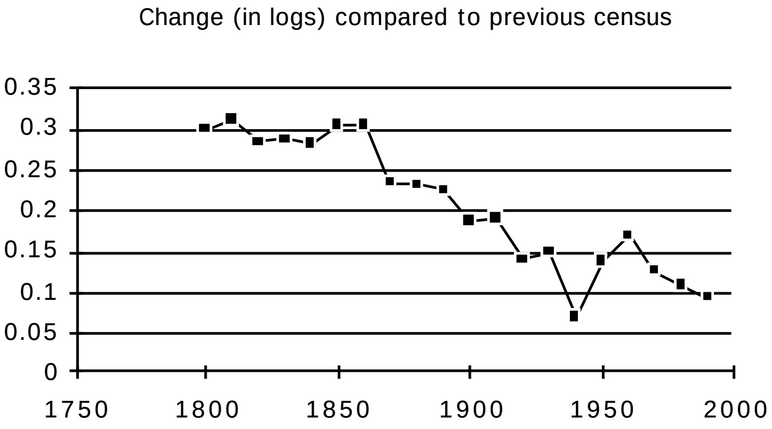
Second, I am worried by the appearance of "cycling" in these residuals. Is this the signal I have just warned myself about? If fitting a line leaves residuals that are quadratic, if fitting a quadratic leaves residuals that are quartic — then you probably should stop and think rather than proceeding ever further through an infinite regress. But that is not what happened here. Here, I fit a quadratic and got residuals with one peak (the middle) and two valleys, thats one more than I would have gotten if I were locked into a polynomial regress. That doesn't prove that I'm not in trouble. But it is re-assuring. These residuals may, in fact, be real — not the product of a misguided analysis.

You should also note that the comparison between these two analyses shows an example of a general rule that appears to be completely counter-intuitive: You look at a graph and see various bumps and curves. The mind seizes on these features, and the scientist within says: “How can I explain that?” The bumps and patterns look real. They are the “that” that needs to be explained. But it isn’t that simple. In most cases the pubps and patterns are bumps and patterns as compared to some base-line expectation. And if you change the base line expectation, this thing in your head which you make appear on the graph, then you changethebumps and patterns. The mind seizes on these alternative feature and the scientist within goes off in another direction.

With these data the first analysis seized on a bend in the curve, circa 1870. On the graph of residuals for the first 100 years, this point for 1870 was approximately 6% below the residual for 1860. It attempted to “explain” that by breaking the data into two batches — corresponding to an idea that the process itself had changed circa 1870. So the analysis said “explain” — wrap a story around — exponential growth at one rate during the first 100 years, a change in the process, and then exponential growth during the remaining years. The raw material calling for this explanation was the bump. The explanation will focus on some relatively sudden shift, circa 1870.

The second analysis compared the whole (not the parts) to a process in which an initial growth rate of about 3% tapered off slowly and smoothly during the ensuing 200 years, down to a growth rate of about 2%. That reference shifts the attention. Now the interesting remaining reatures are the descents into two valleys and the climbs out of them, two valleys separated by about 100 years. The least interesting part of the graph is the middle: the residuals for 1860 through 1910 are all up about 4%, with nothing remarkable demanding attention to 1870. Compared to a basically constant growth rate, modified by a very slow decline (from 3% to 2% in 200 years), the middle of the curve is flat and uninteresting. The hypothesis handles the middle years very nicely, shifting attention to the two valleys, circa 1840 and circa 1940. The point is that the very phenomenon that we think we have to explain is already, in part, a product of the analysis. How do you reduce the subjectivity that is buiklt into what appear to be facts? By extreme skepticism, by constant testing, by enlarging the research to data that can show themselves to be consistent with a hypothesis, thus supporting it, or inconsistent with a hypothesis, thus rejecting it. We also have rules, like choosing parsimony: In this case, the first description requires two straight lines separated by a break while the second description requires one qudratic equation, with no break. The second is a more parsimonious description. I also have a rule of skepticism when I see things that are too regular. In this case the residuals are too symmetrical, displaying mirror symmetry around the middle. That suggests that there may be a better single equation, that the one I've used is wrong, and that these residuals are a mathematical result of the difference between the better equation and the one I've used.

So, what have I established and what can I speculate

The population of the United States increased from 3.9 million to 250 million during the 200 years from 1790 to 1990.	✓ Fact
The rate of increase has declined from about 3 percent per annum to about 1 percent per annum.  	✓ Fact

Relative to these long term trends there have been two short term increases in the rate of increase, one following 1840, the other following 1940.	- Relative statement, true in stated context.
The first increase may be due to massive immigration following the brailroad expansion into farm land of the “West” and the post World War II baby boom.	? Speculation. Really unacceptable speculation, were this a final report, because both statements point to other data that could have been presented but have not been. In the first case, immigration data can be checked to see whether or not it changes circa 1840 and whether or not its magnitude is sufficient to account for the bump in the residuals. In the second case, birth rates can be checked to see whether these birth rates can acccount for the bump and whether changes in the birth rate were sufficient to create the observed change in the rate of increase for the total population (of all ages).

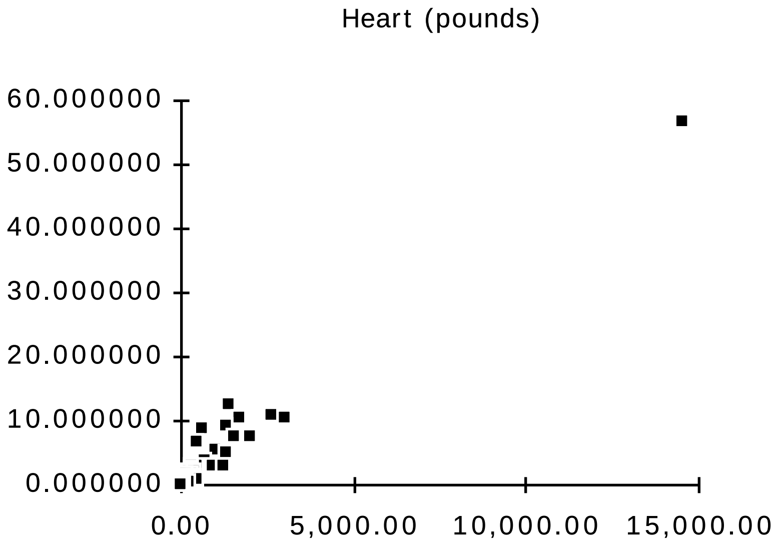
Log Log

"Log log graphs", as they are called are simple, in principle, but they have a strange reputation — among those people with whom such things have any reputation at all.

I think what happens is that people who use statistical tools, and whose control over the techniques they depend on is tenuous, suddenly realize that things have gotten beyond them with log log analysis. It is too many steps beyond control, and they get scared. The discomfort is expressed as skepticism, but there is actually an event that triggers the expression of: It seems to be a common experience that many kinds of things look linear — when you are looking at them on a log log graph. So people, the same people with whom these things have any reputation at all, say "everything" looks linear on a log log graph, and back off. They dismiss what they see — the idea being that if everything looks linear on a log log graph, then nothing is learned in any particular case.

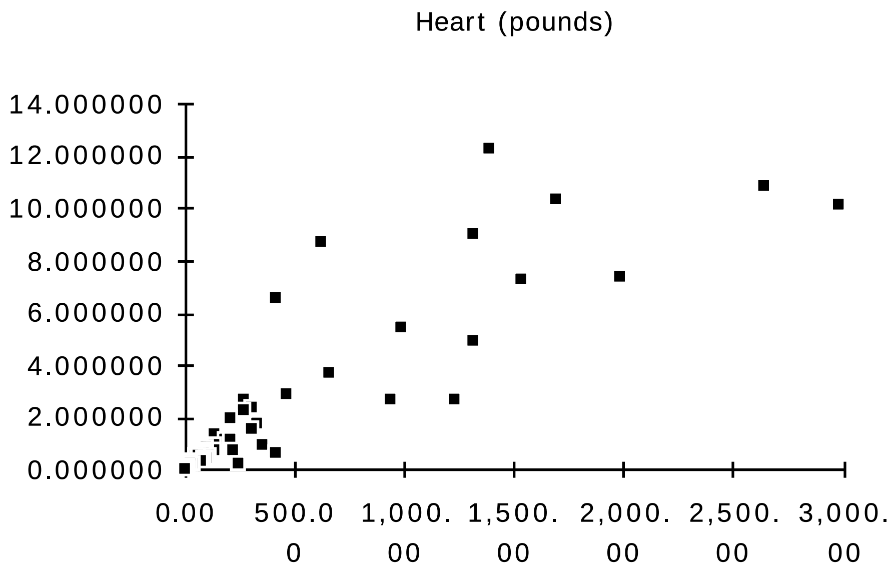
Such stuff somehow passes for sophistication, but really it is a kind of belief in magic. It comes from statistics as magic, that then gets out of control. But if we stay calm and rational, if we use fairly simple math — and believe in it as the tool by which to interpret what you've found, there is nothing out of control when logs begin to pop up on both sides of an equation. Would that it were true that "everything" looks linear on a log log graph. That has not been my experience. And if many things do look linear with this analysis then there is something to be learned here about nature.

First let's take a look: Here, for example is Heart Weight (as "Y" — shown vertically) and Body Weight. The data are from ___ describing the average weights of bodies and of the various organs of vertebrates (a very peculiar collection of data). You might thing about lines on such a graph, and then think about slopes — looking for the weight of the heart as a fraction of the weight of the body, the larger the body the larger the heart. That's what I expect. But all bets are off when you look at the graph of these weights. Looking at the graph, and speaking non-technically, it's a mess.



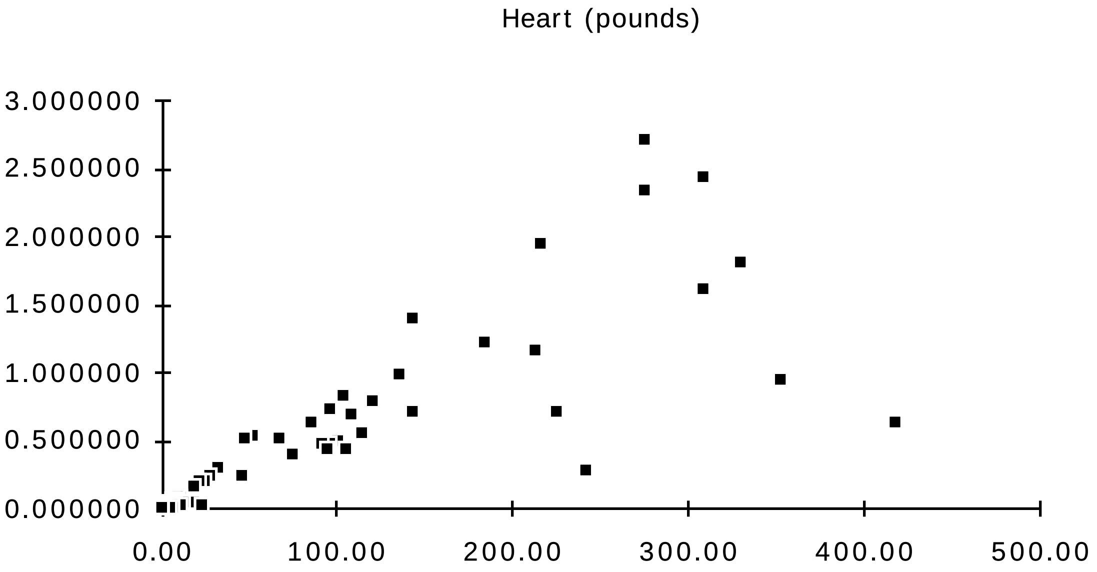
Now, what *you* should do if somebody offers you such a graph, as I have offered it to you is either a), walk away because this person doesn't know how to analyze data, or b) gently walk toward such a person to explain that data analysis begins "at the beginning". And the beginning is one variable analyses — always accompanied by good labels (not a dot at the upper left, but "elephant" (In fact, I would go so far as to say that labels are so important with such things that you will get further faster with these data doing them by hand, given the unlikeliness of getting current software to label properly, by machine.) And there you will discover, or guess, or hypothesize, because you've dealt with such things before, that the intelligent start, for this, would have been logs, not pounds.

But here's what would be done by the pseudo sophisticated. Actually, this person is starting out wrong and then patching and filling — trying to make the patching and filling look like sophistication. Patching and filling, we observe that that point ("Elephant", were it labeled) is an outlier — exclude it from the analysis (or analyze it separately). O.K., excluding the “outlier”



That looks better, until you realize that seeing about 15 points clearly may look better, but that “15” is 15 out of about 150 points, and most of the stuff is still down there at the lower left.

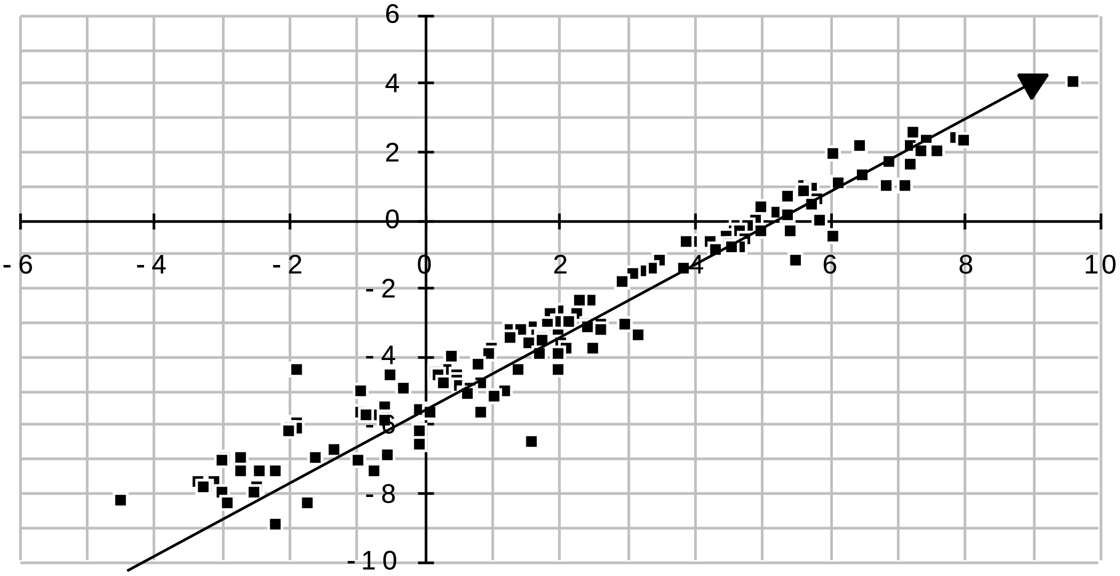
How about excluding just those 15 or so points that *are* the most visible, concentrating on the main body of the data — in effect declaring these 15 as outliers too.

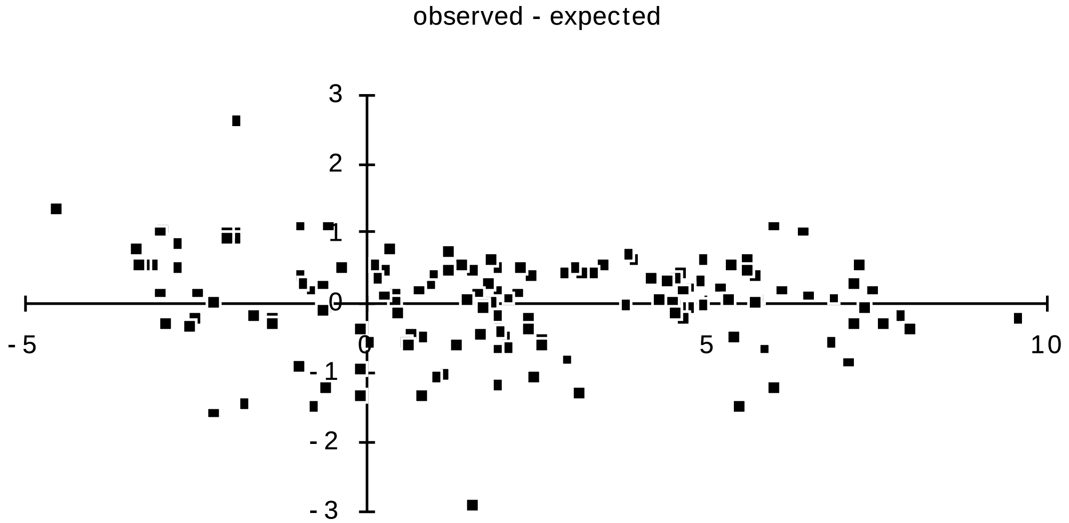


Well — better. But I'd venture to say that if this were not already the third graph in the sequence, if we were starting fresh with it, then we would observe that the 20 or so points that are visible at the upper right seem to be on a different scale from the bulk of the data, about 120 point, at the lower left. And we would proceed to remove them from the data, just as we have already removed 16 points in order to get at the heart of the data.

When you find yourself getting into this kind of dissatisfying loop, cutting out data, cutting out more data, and still not really solving the problem, then it is time to drop back and think. And thinking, or guessing, or starting with the logs in the first place, gets this (using logs base e).

Ln Heart Wt





There you see a line, or at least a sort-of-linear cloud. And, you can understand why someone starting with the first figure (which they should not have) would be in awe of this log log picture derived from the same data: Junk has become orderly and, if you have let the mathematics get the better of you, then the conversion of the chaos in the first figure into the order of the second figure, might seem like magic. Magic had nothing to do with it.

Below it, observing the line and then looking at residuals, I've computed residuals using the numbers

$$\ln\left(Heart\right) = .9845 \ \ \ln\left(Body\right) - 5.15925$$

So, if we're so smart, what does this log log equation mean? I find out by trusting the math and using it to decode the meaning. First exponentiate

$$e^{\ln(Heart)} = e^{.9845 \ \ln(Body) - 5.15925}$$

$$e^{\ln(Heart)} = e^{(.9845 \ \ln(Body) - 5.15925)}$$

and then simplify

$$Heart = e^{-5.15925} \ Body^{.9845}$$

$$Heart = e^{(-5.15925)} e^{(.9845 \ \ln(Body))}$$

and

$$Heart = .005746 \bullet Weight^{.9845}$$

$$Heart = .005746 \ Weight^{.9845}$$

It says: The weight of the heart, on the left, is proportional to the .98th power of the weight of the organism. Observing that .9845 is close to one, the equation says that the weight of the heart is proportional to the weight of the body.

$$Heart \approx .005746 \ Body$$

So the equation implies that the weight of the heart is to the weight of the body as .005746.

$$\frac{Heart}{Body} \approx .005746$$

That is, weight of the heart is approximately one half of one percent of the weight of the body (0.57 %). That is the decoding of the slope and the intercept of this log log linear relation.

Thinking about this relation, I find the relation surprising: It says that the weight of the heart is *directly proportional* to the weight of the body. I'm not sure exactly what I expected (so much for falsifiable hypotheses), but I was thinking that big animals tend to be warm blooded, that should make a difference in demands on heart. Or, there should be some efficiencies of size when a heart muscle has to push blood around. Too bad: what differences there are in the residuals, they are not a function of total weight.

How big are the residuals?

You can read that right off of the graph of the residuals: The observed values are within plus or minus one of the predicted values. (Imagine trying to estimate the size of the residuals using the first graph, without logs). And errors of plus or minus one in logs, base e, correspond to factors of 2.7 above or below the predicted weight of the heart. So

The weight of the heart is proportional the weight of the body, averaging about one half of one percent of body weight. However, there is a large variation around that average, amounting to a factor of nearly three in either direction, meaning that heart weight typically falls within a range of 0.2 percent to 1.6 percent of body weight (multiplying and dividing .005746 by 2.7).

Efficiencies and Inefficiencies of Scale

Those of you with an unbreakable tendency to stay up late, with the television talking to itself in the other room, have seen the monster films of the ‘50’s. Well, whatever your concerns over monster bugs, worms grown large enough to devour cities, and earth pounding giants, I can assure you there is no need to worry. It’s not just that I’ve looked around and found no footprints in the mud, it is that some things really are impossible.

Consider the case of the giant worm. Now worms, even those with aggressive tendencies have certain base line tasks to handle, like eating and breathing. If you are a worm then the way you breath is through that nice moist skin. This is where oxygen is absorbed from the outside and where other gases are released.

Now suppose that a giant film director takes charge of your wormly fate and finds some way of making you twice as large, not by moving the camera in to half the distance, but really making you twice as large. Here’s what happens: You become twice as long. You are twice as wide, you are twice as high. That means that you weight two times two times two times as much as you did before. Twice as big implies eight times as much meat on the body. That means you have to take in about eight times the oxygen and release eight times the waste.

So, you’ve got to process about eight times as much air as before and, being a worm, you are going to do it through your skin. How much skin have you got to work with? Well, allowing that the cross section of a worm is roughly circular, your diameter increased by a factor of 2 so your circumference has increased by a factor of 2. Allowing that you are now twice as long, your skin surface will have increased four fold: Your mass (in need of oxygen) has increased eight times fold but your skin has increased only four fold.

Now, of course, increasing the diameter of an earth worm from one eighth of an inch to one quarter of an inch would be quite an event in the life of the worm, but it would still be nothing in the competition among midnight film monsters. So, suppose the worm is be ten times larger than normal: A one meter long earth worm should create at least a little alarm. This worm, is now ten times wider, ten times higher, and ten times longer. He has 1,000 times the meat, but only 100 times the skin. The ratio of meat to air exchanging skin has diminished ten fold. "Breathing" is going to be a problem.

Do that to the worm in one generation, by a horror induced mutation and it will die, literally, of its own weight long before it has terrified too many citizens. However, given a few tens of millions of years to accommodate to this transition, the worm might evolve a more convoluted skin, increasing its surface. It might protect all that convoluted moist skin by some sort of protective structure. And it might develop a mechanism for drawing the outside air over this protected convoluted moist skin — which might end up looking pretty much like a lung.

The film of this evening features a giant ant, born in the Nevada deserts (near the test sites) which, according to the film's scientist (he's wearing the white coat) is "two meters in length, that's more than nine feet". Bad start. Whatever: He is roughly 400 times the length of your garden variety ant. That's 64,000,000 times the weight and he's got to carry that weight on legs that are probably still a bit spindly at about an inch. Since the carrying capacity of a column (his leg) is proportional to its cross section, this guy is going to need leg splints, or some leg transplants from a hippopotamus.

The constraints by which form and function become related are referred to as "allometry", the study of form. (D'Arcy Wentworth Thompson citation ***) In biology (and engineering) there are compelling laws by which the extension of an object, the surface of the object, and the mass of the object are necessarily closely related in non-linear relations.

In the more social of the sciences, the principles of form create questions. If the dense population centers of San Francisco were to spread over more of the ground (in two dimensions) and perhaps go high rise (in a third dimension), and if the flow through traffic arteries had to keep pace by creating wider and wider freeways (in one dimension), then you would have to pave the peninsula to supply transportation. Maybe.

Or consider, what is the ratio (or, perhaps, the optimal ratio) of administrators to students or administrators to faculty in a University? How is it related to size? I would think that if you took one well run university of 5,000 students and ran it up to 10,000 students then, at the worst, the ratio of administrators to students would remain constant: After all, the worst case would be simply running the 10,000 student entity as two 5,000 student universities, doubling everything and leaving t he ratios intact. Anything better would be an economy of scale. Anything less would be a pathology of scale.

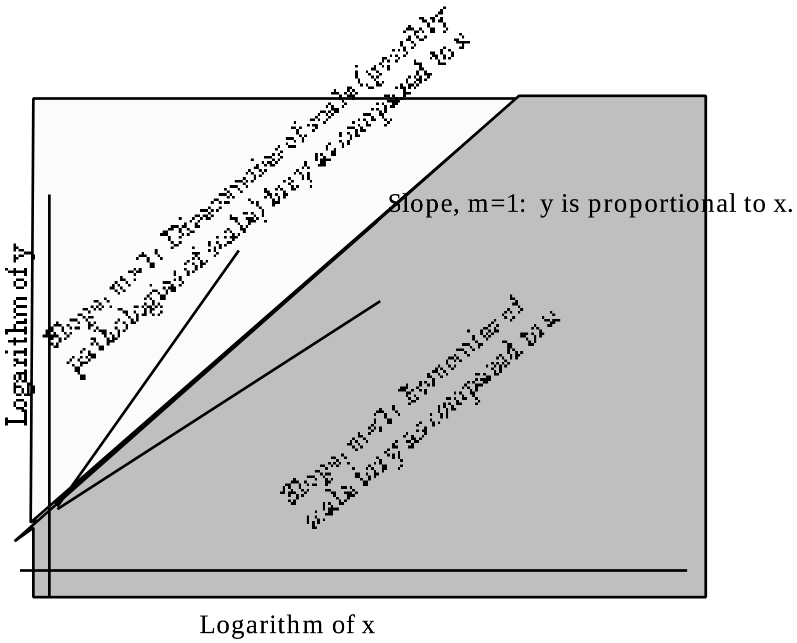
Consider: Does a ten fold difference in the size of two steel companies correspond to a ten fold increase in the number of employees or a ten fold increase in its capital assets? (Is there a change in the production function related to size?) Does a ten fold difference in the size of two cities correspond to a ten fold difference in the cost / taxes / police force / roads ? Is it more or less than ten fold? Does a ten fold difference in the size of two nations correspond to a ten fold increase in the sizes of their military forces? More or less?

When you hear a question framed in the form does an a-fold increase in this correspond to a b-fold increase in that, you have entered the domain of log log graphs.

Where log(y) = m log(x)+b the easy case is m=1 (at least approximately). Then y is directly proportional to x. At m=1, the slope in the log log curve establishes the direct proportionality. The intercept establishes the proportion. And the size of the residuals works exactly the way it did in the semi-log graph: The residuals are residuals with respect to log y and they specify the factors by which the residuals lie above or below the predicted values.

Where log(y) = m log(x)+b with m>1, comparing one case to another differences in y variable are proportionately greater than differences in the x variable. The slope establishes the relation. E.g., m=2 implies that differences in y correspond to differences in the square of x. The intercept establishes the comparison at one reference point, log x = 0, x =1. And the residuals again specify the factors by which the residuals lie above or below the predicted values.

Where log(y) = m log (x)+b with m<1, differe3nces in the y variable are proportionately less than differences in the x variable. For example, if you were comparing the number of employees to the dollar assets of a steel company, for a wide range of steel companies, then with y = assets, m>1 would tell you that large steel companies are more asset/capital intensive than small steel companies. m<1 would tell you that large steel companies are more labor intensive than small steel companies. And m=1 would tell you that the production function governing the mix between capital and labor was unrelated to size.


Slope: m>1: Diseconomies of scale
Slope, m=1: y is proportional to x.
Logarithm of y
Slope: m<1: Economies of
Logarithm of x

Exercises:

Describe the relation between the weight of the body and the weight of the lungs as implied by Spector’s data.

Exercises:

Describe the relation between the weight of the body and the weight of the lungs as implied by Spector's data.

#	Species	Sex & Number	Body Wt. KG	ln(Body Wt)	Brain grams	Ln(Brain)	Heart Grams	LN(Heart grams)	Liver Grams	LN(Liver)	Lungs Grams	Ln(Lungs)	Weight (pounds @2.2046 pounds per kilogram)
1	Man (Homo sapiens) Australian aborigine	M1	76	4.331	1345.2	7.204	0	????	0	????	0	????	167.55
2	Man (Homo sapiens) Chinese	M1	84	4.431	1478.4	7.299	554.4	6.318	2041.2	7.621	0	????	185.19
3	Man (Homo sapiens) Filipino	M1	43	3.761	1105.1	7.008	197.8	5.287	0	????	0	????	94.80
4	Man (Homo sapiens) Indian, Maya Quiche	M1	42	3.738	1268.4	7.146	218.4	5.386	1041.6	6.949	1314.6	7.181	92.59

5	Man (Homo sapiens) Indian, Maya Quiche	F1	46	3.829	1002.8	6.911	225.4	5.418	0	????	0	????	101.41
6	Man (Homo sapiens) Negro	F7	47	3.850	1283.1	7.157	380.7	5.942	1320.7	7.186	0	????	103.62
7	Man (Homo sapiens) White, American	F4	49	3.892	1239.7	7.123	313.6	5.748	1127	7.027	357.7	5.880	108.03
8	Man (Homo sapiens) White, European	F4	49	3.892	1239.7	7.123	313.6	5.748	0	????	0	????	108.03
9	Agouti (Dasyprocta punctata)	FM5	2.6	0.956	15.08	2.713	13.26	2.585	73.84	4.302	5.72	1.744	5.73
10	Antbear (Cyclops didactylus)	?1	0.09	-2.408	4.293	1.457	0	????	0	????	0	????	0.20
11	Anteater (tamanduas tetradactyla)	MF4	2.2	0.788	23.98	3.177	0.66	-0.416	58.08	4.062	23.1	3.140	4.85

12	Armadillo (Dasypus novemcinctus)	MF12	3.3	1.194	8.25	2.110	9.24	2.224	0	????	23.1	3.140	7.28
13	Ass (Equus asinus)	F1	150	5.011	405	6.004	825	6.715	1260	7.139	1245	7.127	330.69
14	Bat. vampire (Desmodus rotundus)	MF5	0.028	-3.576	0.9352	-0.067	0	????	0	????	0	????	0.06
15	Bear, brown (Ursus americanus)	F1	550	6.310	0	????	0	????	0	????	0	????	1,212.53
16	Bear, grizzly (U. horribilis)	F1	140	4.942	224	5.412	1106	7.009	0	????	0	????	308.64
17	Beaver (Castor canadensis)	M1F1	5	1.609	22.5	3.114	21.5	3.068	151.5	5.021	48.5	3.882	11.02
18	Bison, American (Bison bison)	F1	55	4.007	335.5	5.816	363	5.894	698.5	6.549	1193.5	7.085	121.25
19	Buffalo, African (Syncerus caffer)	M3F1	700	6.551	630	6.446	3290	8.099	6860	8.833	6580	8.792	1,543.22

20	Bushbok (Tragelaphus scriptus)	M1F1	44	3.784	162.8	5.093	334.4	5.812	858	6.755	721.6	6.581	97.00
21	Camel, bactrian (Camelus bactrianus)	M1	450	6.109	540	6.292	0	????	0	????	0	????	992.07
22	Caribou, ground (Rangifer arcticus)	M3F1	98	4.585	294	5.684	882	6.782	1793.4	7.492	2058	7.629	216.05
23	Cat,domestic (Felis catus)	M7F3	3.3	1.194	25.41	3.235	14.85	2.698	118.47	4.775	34.32	3.536	7.28
24	Cattle, Holstein (Bos taurus)	M5	900	6.802	450	6.109	3330	8.111	8280	9.022	6210	8.734	1,984.14
25	Cattle, Holstein (B. taurus)	F198	600	6.397	420	6.040	2220	7.705	7200	8.882	4320	8.371	1,322.76
26	Cheetah (Acinonyx jubatus)	F2	21	3.045	81.9	4.405	107.1	4.674	676.2	6.516	243.6	5.496	46.30
27	Chimpanzee (Pan troglodytes)	M1	52	3.951	436.8	6.079	249.6	5.520	0	????	0	????	114.64

28	Chimpanzee (P. troglogytes)	F1	44	3.784	325.6	5.786	220	5.394	1210	7.098	598.4	6.394	97.00
29	Chipmunk (Tamias striatus)	F2	0.07	-2.659	2.072	0.729	5.572	1.718	5.18	1.645	0.672	-0.397	0.15
30	Coati (Nasua nasua)	M2	5.1	1.629	33.66	3.516	19.38	2.964	83.13	4.420	23.97	3.177	11.24
31	Coyote (Canis latrans)	F2	8.5	2.140	0	????	72.25	4.280	292.4	5.678	61.2	4.114	18.74
32	Deer, white-tailed (Odocoileus viginianus)	M1	65	4.174	208	5.338	630.5	6.447	1020.5	6.928	0	????	143.30
33	Dog (Canis familiaris)	M2F2	13	2.565	76.7	4.340	110.5	4.705	382.2	5.946	122.2	4.806	28.66
34	Elephant (Loxondonta africana)	M1	6600	8.795	5280	8.572	25740	10.156	106920	11.580	137280	11.830	14,550.36
35	Fox, gray (Urocyon cineroargeneus)	M1	3.8	1.335	37.62	3.628	22.04	3.093	51.3	3.938	19.38	2.964	8.38
36	Fox, red (Vulpes fulva)	F1	4.6	1.526	52.9	3.968	41.4	3.723	0	????	0	????	10.14

37	Gazelle (Gazella thomsoni)	M2	24	3.178	91.2	4.513	240	5.481	516	6.246	276	5.620	52.91
38	Giraffe (Giraffa camelopardalis)	F1	1200	7.090	720	6.579	4920	8.501	18720	9.837	11880	9.383	2,645.52
39	Goat (Capra hircus)	F1	28	3.332	114.8	4.743	0	????	532	6.277	0	????	61.73
40	Gorilla (Gorilla gorilla)	M1	180	5.193	0	????	0	????	0	????	0	????	396.83
41	Guinea pig (Cavia porcellus)	M58	0.26	-1.347	3.458	1.241	1.378	0.321	13.364	2.593	3.068	1.121	0.57
42	Guinea pig (C. porcellus)	F10	0.43	-0.844	3.956	1.375	1.677	0.517	16.598	2.809	4.601	1.526	0.95
43	Hamster, golden (Mesocricetus auratus)	M2F2	0.12	-2.120	1.056	0.054	0.564	-0.573	6.192	1.823	0.552	-0.594	0.26
44	Hare, African (Lepus capensis)	F1	2.9	1.065	10.15	2.317	29.58	3.387	51.33	3.938	17.69	2.873	6.39

45	Hippopotamus (Hippopotamus amphibius)	F1	1350	7.208	675	6.515	4590	8.432	23625	10.070	11340	9.336	2,976.21
46	Horse, Percheron (Equus caballus)	M1	635	6.454	635	6.454	5588	8.628	8509	9.049	5715	8.651	1,399.92
47	Horse, Percheron (E. caballus)	F1	770	6.646	616	6.423	4697	8.455	6699	8.810	5390	8.592	1,697.54
48	Hyena, spotted (Crocuta crocuta)	M2	62	4.127	173.6	5.157	446.4	6.101	3174.4	8.063	6770.4	8.820	136.69
49	Hyrax (Heterohyrax brucei)	M1	0.75	-0.288	12.3	2.510	3.6	1.281	31.5	3.450	5.55	1.714	1.65
50	Jackal (Canis mesomelas)	M2	2.8	1.030	45.08	3.808	21	3.045	120.4	4.791	29.4	3.381	6.17
51	Jaguar (Felis onca)	F1	34	3.526	146.2	4.985	183.6	5.213	880.6	6.781	567.8	6.342	74.96
52	Kinkajou (Potos flavus)	F1	2.6	0.956	30.68	3.424	14.04	2.642	97.76	4.583	77.74	4.353	5.73

53	Lemming, rock (Dicrostonyx rubricatus)	M4	0.05	-2.996	0.085	-2.465	0.295	-1.221	2.525	0.926	0.795	-0.229	0.11
54	Leopard (Panthera pardus)	M1	48	3.871	134.4	4.901	201.6	5.306	897.6	6.800	499.2	6.213	105.82
55	Lion (P. leo)	M4	125	4.828	237.5	5.470	1062.5	6.968	0	????	2650	7.882	275.58
56	Lion (P. leo)	F3	97	4.575	194	5.268	523.8	6.261	3142.8	8.053	1998.2	7.600	213.85
57	Lynx (Lynx baileyi)	M1	7.4	2.001	0	????	0	????	0	????	0	????	16.31
58	Manatee (Trichechus manatus)	M1	425	6.052	340	5.829	1232.5	7.117	5525	8.617	3060	8.026	936.96
59	Manatee (T. manatus)	F1	560	6.328	0	????	1232	7.116	6272	8.744	3752	8.230	1,234.58
60	Mole (Scalopus aquaticus)	M1	0.04	-3.219	1.172	0.159	0.276	-1.287	1.564	0.447	0.744	-0.296	0.09
61	Mongoose (Ichneumia albicauda)	M1	4.4	1.482	28.16	3.338	28.16	3.338	61.16	4.113	58.08	4.062	9.70

62	Monkey, blackhowler (Alouatta palliata)	MF28	6.2	1.825	50.22	3.916	20.46	3.018	201.5	5.306	39.06	3.665	13.67
63	Monkey, rhesus (Macaca mulatta)	M4	3.3	1.194	91.74	4.519	12.54	2.529	68.97	4.234	0	????	7.28
64	Monkey, rhesus (M. mulatta)	F7	3.6	1.281	92.52	4.527	12.24	2.505	0	????	68.04	4.220	7.94
65	Mouse, jumping (Zapus hudsonicus)	M1F3	0.018	-4.017	0.6426	-0.442	0.1854	-1.685	1.0134	0.013	0.2412	-1.422	0.04
66	Mouse, meadow (Microtus drummondi)	MF67	0.023	-3.772	0.0667	-2.708	0.1564	-1.855	1.0488	0.048	0.391	-0.939	0.05
67	Muskrat (Ondatra zibethica)	M1	0.9	-0.105	5.31	1.670	3.24	1.176	21.96	3.089	4.32	1.463	1.98
68	Opossum, woolly (Philander laniger)	M1F1	190	5.247	0	????	3002	8.007	9006	9.106	3002	8.007	418.87

69	Porcupine (Erethizon dorsatum)	M1F3	2.9	1.065	22.62	3.119	15.95	2.769	116	4.754	28.42	3.347	6.39
70	Porpoise (Phocaena phocaena)	M1	140	4.942	1708	7.443	728	6.590	2912	7.977	5166	8.550	308.64
71	Rabbit, giant Flemish (Lepus spp)	M2	3.7	1.308	10.73	2.373	10.73	2.373	98.42	4.589	0	????	8.16
72	Rabbit, giant Flemish (Lepus app)	F22	2.5	0.916	10	2.303	8.75	2.169	79.75	4.379	13.25	2.584	5.51
73	Raccoon (Procyon lotor)	M1	5.2	1.649	42.64	3.753	42.12	3.741	186.16	5.227	186.16	5.227	11.46
74	Raccoon (P. lotor)	F1	2.2	0.788	33.22	3.503	19.58	2.975	138.38	4.930	19.14	2.952	4.85
75	Rat, Norway (Rattus norvegicus)	M2F1	0.25	-1.386	3.05	1.115	1.3	0.262	8.375	2.125	1.975	0.681	0.55
76	Reedbuck (Redunca redunca)	M2	31	3.434	105.4	4.658	235.6	5.462	511.5	6.237	415.4	6.029	68.34
77	Seal, ringed (Phoca hispida)	M3F2	39	3.664	245.7	5.504	284.7	5.651	1095.9	6.999	721.5	6.581	85.98

78	Shrew (Blarina brevicauda)	M29	0.02	-3.912	0.374	-0.983	0.204	-1.590	1.162	0.150	0.448	-0.803	0.04
79	Shrew (Blarina brevicauda)	F39	0.017	-4.075	0.3587	-1.025	0.1785	-1.723	0.9265	-0.076	0.3723	-0.988	0.04
80	Skunk (Mephitis mephitis)	M1F2	2.1	0.742	6.93	1.936	12.18	2.500	56.49	4.034	33.39	3.508	4.63
81	Sloth (three-toed (Bradypus tridactylus)	MF6	1.8	0.588	13.5	2.603	0	????	0	????	0	????	3.97
82	Squirrel, red (sciurus hudsonicus)	M4	0.18	-1.715	4.626	1.532	1.548	0.437	3.924	1.367	2.61	0.959	0.40
83	Squirrel, red (s. husdonicus)	F4	0.25	-1.386	5.05	1.619	1.825	0.602	6.7	1.902	3.2	1.163	0.55
84	Steinbok (Raphicerus campestris)	M2	8.6	2.152	49.02	3.892	72.24	4.280	174.58	5.162	149.64	5.008	18.96
85	Swine (Sus scrofa)	F36	102	4.625	0	????	326.4	5.788	1540.2	7.340	0	????	224.87
86	Tapir (Tapirella bairdii)	M1F1	11.4	2.434	0	????	96.9	4.574	349.98	5.858	239.4	5.478	25.13

87	Tiger (Panthera tigris)	F1	160	5.075	224	5.412	432	6.068	1824	7.509	1024	6.931	352.74
88	Walrus (Odobenus rosmarus)	M1F3	600	6.397	1020	6.928	4080	8.314	17520	9.771	8160	9.007	1,322.76
89	Warthog (Phacochoerus aethiopicus)	M1	65	4.174	123.5	4.816	325	5.784	1495	7.310	546	6.303	143.30
90	Weasel, arctic (Mustela arctica)	M3F1	0.18	-1.715	5.04	1.617	3.078	1.124	8.532	2.144	3.744	1.320	0.40
91	Whale, white (Delphiapterus leucas)	M4	447	6.103	2324.4	7.751	2458.5	7.807	6794.4	8.824	12069	9.398	985.46
92	Whale, white (Delphiapterus leucas)	F2	300	5.704	2340	7.758	1710	7.444	4770	8.470	7860	8.970	661.38
93	Wildebeest (Connochaetes taurinus)	M2	210	5.347	441	6.089	1302	7.172	2247	7.717	2814	7.942	462.97
94	Wolf (Canis lupus)	M1	22	3.091	114.4	4.740	237.6	5.471	607.2	6.409	783.2	6.663	48.50

95	Zebra (Equus quagga)	M3F1	280	5.635	560	6.328	3976	8.288	4676	8.450	2240	7.714	617.29
96	Blackbird (Quiscalus quiscala)	F1	0.08	-2.526	2.848	1.047	0.112	-2.189	2.568	0.943	0.168	-1.784	0.18
97	Bluebird (Sialia sialis)	M1F1	0.03	-3.507	1.272	0.241	0.417	-0.875	0	????	0	????	0.07
98	Buzzard, steppe (Buteo vulpinus)	M1	0.56	-0.580	7.896	2.066	4.592	1.524	10.864	2.385	4.648	1.536	1.23
99	Catbird (Dumatella carolinensis)	F1	0.03	-3.507	0.129	-2.048	0.297	-1.214	0	????	0.552	-0.594	0.07
100	Canary (Serinus canarius)	M1F1	0.016	-4.135	0.7552	-0.281	0.2064	-1.578	0.8624	-0.148	0.024	-3.730	0.04
101	Cowbird (Molothrus ater)	F1	0.07	-2.659	2.856	1.049	1.127	0.120	0	????	0	????	0.15
102	Crane, gray (Grus canadensis)	M1	1.6	0.470	8.32	2.119	18.4	2.912	28.48	3.349	14.88	2.700	3.53

103	Crow (Corvus brachyrhyncos)	M1	0.33	-1.109	9.108	2.209	3.135	1.143	0	????	9.768	2.279	0.73
104	Duck, pintail (Anas acuta)	F1	0.67	-0.400	4.958	1.601	8.308	2.117	30.351	3.413	17.152	2.842	1.48
105	Eagle, tawny (Aguila rapax)	M2F3	2.4	0.875	14.16	2.650	15.12	2.716	43.68	3.777	24.96	3.217	5.29
106	Egret, great white (Casmerodius albus)	F1	10	2.303	59	4.078	90	4.500	320	5.768	321	5.771	22.05
107	Flamingo (Phoeniconaias minor)	M3F2	15	2.708	73.5	4.297	141	4.949	402	5.996	220.5	5.396	33.07
108	Fowl, domestic (Gallus domesticus)	M8	0.73	-0.315	2.92	1.072	4.161	1.426	16.133	2.781	4.38	1.477	1.61
109	Fowl, domestic (Gallus domesticus)	F16	0.61	-0.494	2.684	0.987	3.843	1.346	14.396	2.667	3.721	1.314	1.34

110	Fowl, white leghorn, "germ-free"	?	.9-1.2	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
111	Goose, Egyptian (Alopochen aegypticus)	F1	1.9	0.642	7.41	2.003	18.24	2.904	33.63	3.515	34.2	3.532	4.19
112	Guineafowl (Numida meleagris)	M1	1.6	0.470	4.16	1.426	14.08	2.645	28.16	3.338	28.64	3.355	3.53
113	Gull, herring (Larus argentatus)	F2	0.53	-0.635	5.035	1.616	5.194	1.648	27.136	3.301	0	????	1.17
114	Hawk, red-tailed (Buteo borealis)	F3	1	0.000	9.7	2.272	6.7	1.902	13.7	2.617	9	2.197	2.20
115	Hummingbird (Amazilia tzacatl)	F1	0.005	-5.298	0.208	-1.570	0.1185	-2.133	0.2615	-1.341	0.01	-4.605	0.01
116	Ostrich, masai (Struthio camelus)	M1	125	4.828	37.5	3.624	1225	7.111	2075	7.638	2950	7.990	275.58

117	Owl, honed (Buteo viginianus)	M1	1.2	0.182	13.92	2.633	8.76	2.170	0	????	10.92	2.391	2.65
118	Partridge (Francolinus sephaena)	M1	0.21	-1.561	1.512	0.413	1.47	0.385	8.736	2.167	0	????	0.46
119	Pelican (Pelecanus occidentalis)	F2	3.3	1.194	17.82	2.880	22.11	3.096	73.26	4.294	30.03	3.402	7.28
120	Pheasant (Phasianus cochicus)	M1	0.62	-0.478	3.286	1.190	5.58	1.719	9.052	2.203	0	????	1.37
121	Pigeon (Columba livia)	M3F1	0.27	-1.309	2.565	0.942	4.725	1.553	4.752	1.559	0	????	0.60
122	Raven (Corvus corax)	F1	1.25	0.223	35.125	3.559	10.625	2.363	0	????	0	????	2.76
123	Robin (Turdus migratorius)	M2	0.07	-2.659	2.107	0.745	1.022	0.022	0	????	1.694	0.527	0.15
124	Sparrow (Passer domesticus)	M75	0.024	-3.730	1.0464	0.045	0.4152	-0.879	1.2288	0.206	0.3744	-0.982	0.05

125	Sparrow (Passer domesticus)	F11	0.023	-3.772	1.0074	0.007	0.3887	-0.945	1.0741	0.071	0.3956	-0.927	0.05
126	Starling (Sturnus vulgaris)	M15	0.06	-2.813	1.956	0.671	0.972	-0.028	2.076	0.730	1.122	0.115	0.13
127	Starling (Sturnus vulgaris)	F10	0.06	-2.813	1.878	0.630	0.894	-0.112	2.256	0.814	1.122	0.115	0.13
128	Stork, European (Ciconia ciconia)	M2F1	3.3	1.194	15.51	2.741	30.36	3.413	63.36	4.149	36.63	3.601	7.28
129	Alligator (Alligator mississipiensis)	M2	190	5.247	13.3	2.588	285	5.652	722	6.582	1026	6.933	418.87
130	Crocodile (Crocodylus acutus)	M1F1	110	4.700	11	2.398	132	4.883	1122	7.023	1100	7.003	242.51
131	Iguana lizard (Iguana iguana)	F1	1.3	0.262	0	????	2.47	0.904	32.37	3.477	3.64	1.292	2.87
132	Lizard (Lacerta viridis)	MF15	0.05	-2.996	0.12	-2.120	0.06	-2.813	2.5	0.916	0	????	0.11

133	Snake, black (Coluber constrictor)	M1F2	0.43	-0.844	0.301	-1.201	0.946	-0.056	2.58	0.948	3.44	1.235	0.95
134	Snake, boa (Boa imperator)	F1	1.8	0.588	0.36	-1.022	5.58	1.719	29.88	3.397	13.68	2.616	3.97
135	Snake, green (Zamenis viridis)	M3F3	0.022	-3.817	0.209	-1.565	0	????	0.4818	-0.730	0	????	0.05
136	Snake, python (Python molurus)	M1	6.1	1.808	1.22	0.199	18.3	2.907	0	????	0	????	13.45
137	Snake, watermoccasin (Ancistrododon pisci)	F1	0.73	-0.315	0.657	-0.420	4.745	1.557	64.605	4.168	22.776	3.126	1.61
138	Toad, horned (Phrynosoma cornutum)	M2F3	0.025	-3.689	0.13	-2.040	0.11	-2.207	0	????	0	????	0.06
139	Turtle (Aromochelys tristycha)	M1	0.12	-2.120	0	????	0.516	-0.662	3.36	1.212	1.02	0.020	0.26

140	Turtle (Aromochelys tristycha)	F2	0.09	-2.408	0	????	0.432	-0.839	2.61	0.959	0.684	-0.380	0.20
141	Turtle (Testudo graeca)	MF30	0.32	-1.139	0.288	-1.245	0	????	8.512	2.141	0	????	0.71
142	Turtle, cumberland (Chrysemys elegans)	M21	0.84	-0.174	0	????	2.688	0.989	45.612	3.820	8.988	2.196	1.85
143	Turtle, cumberland (Chrysemys elegans)	F1	0.86	-0.151	0	????	2.666	0.981	50.912	3.930	7.224	1.977	1.90
144	Frog, bull (Rana catesbiana)	M7	0.49	-0.713	4.557	1.517	1.568	0.450	13.475	2.601	2.597	0.954	1.08
145	Frog, leopard (R. pipiens)	M10	0.036	-3.324	0	????	0.1548	-1.866	1.0116	0.012	0.306	-1.184	0.08
146	Frog, leopard (R. pipiens)	F19	0.038	-3.270	0	????	0.1824	-1.702	1.0944	0.090	0.2888	-1.242	0.08
147	Barracuda (Sphyraena barracuda)	M3F3	8.8	2.175	3.52	1.258	21.12	3.050	60.72	4.106	0	????	19.40

148	Carp (Cyprinus carpio)	M2F4	1.05	0.049	1.26	0.231	1.575	0.454	0	????	0	????	2.31
149	Codfish (Gadus morrhua)	F1	10.6	2.361	5.3	1.668	15.9	2.766	161.12	5.082	0	????	23.37
150	Haddock (G. aeglefinus)	F6	3.3	1.194	1.98	0.683	5.61	1.725	133.65	4.895	0	????	7.28
151	Mackerel (Scomber vernalis)	M1	0.76	-0.274	0.608	-0.498	0	????	0	????	0	????	1.68
152	Mackerel (Scomber vernalis)	F2	1.5	0.405	1.65	0.501	3	1.099	6.45	1.864	0	????	3.31
153	Perch (Perca flavescens)	M6	0.17	-1.772	0.255	-1.366	0.391	-0.939	1.496	0.403	0	????	0.37
154	Perch (Perca flavescens)	F1	0.19	-1.661	0.323	-1.130	1.463	0.380	2.926	1.074	0	????	0.42
155	Pike (Esox lucius)	M4F3	0.42	-0.868	0.504	-0.685	0.63	-0.462	3.612	1.284	0	????	0.93
156	Salmon (Salmo salar)	M3	3.4	1.224	1.02	0.020	12.24	2.505	68.68	4.229	0	????	7.50

157	Salmon (Salmo salar)	F5	5.4	1.686	1.08	0.077	10.26	2.328	93.42	4.537	0	????	11.90
158	Trout, rainbow (Salmo irideus)	M2	0.26	-1.347	0.442	-0.816	0.442	-0.816	2.574	0.945	0	????	0.57
159	Trout, rainbow (Salmo irideus)	F4	0.22	-1.514	0.418	-0.872	0.286	-1.252	2.178	0.778	0	????	0.49

Describe the relation between gross national product and military expenditures (1975 data). (World Handbook of Social and Political Indicaqtors.***)

"COUNTRY NAME"	"TOT DEFENSE EXP,75"	"GROSS NAT PRODUCT,75"	ln(GNP)	ln(Defense)
AFGN	37	2060	7.630	3.611
ALBN	131	1220	7.107	4.875
ALGR	302	13680	9.524	5.710
ANGL	97	2030	7.616	4.575
ARGN	860	39330	10.580	6.757
AUSL	2480	77010	11.252	7.816
AUST	357	36650	10.509	5.878

BHMS		630		
BHRN	14	580	6.363	2.639
BHTN		80		
BLGM	1720	61470	11.026	7.450
BLGR	1680	18420	9.821	7.427
BNGL	76	7280	8.893	4.331
BNIN	6	390	5.966	1.792
BOLV	59	2040	7.621	4.078
BRBD	1	350	5.858	0.000
BRMA	171	3320	8.108	5.142
BRND	9	410	6.016	2.197
BRZL	2440	110130	11.609	7.800
BTSN	0	230		
CAFR	8	390	5.966	2.079
CHAD	23	460	6.131	3.135
CHLE	331	10130	9.223	5.802
CHNA	32800	315250	12.661	10.398
CLMB	165	13630	9.520	5.106
CMRN	34	2050	7.626	3.526
CMRS		70		
CNDA	3160	158100	11.971	8.058
CNGO	26	670	6.507	3.258
CRCA	0	1890		
CUBA	393	7460	8.917	5.974

CVRD		80		
CYPR	21	780	6.659	3.045
CZCH	3180	53450	10.887	8.065
DMNR	46	3390	8.129	3.829
DNMK	858	34450	10.447	6.755
ECDR	75	4180	8.338	4.317
EGPT	1340	9540	9.163	7.200
ELSL	21	1830	7.512	3.045
EQGN	5	100	4.605	1.609
ETHP	110	2730	7.912	4.700
FIJI	1	620	6.430	0.000
FNLD	348	25520	10.147	5.852
FRG	14700	412480	12.930	9.596
FRNC	11400	314080	12.657	9.341
GBON	14	1360	7.215	2.639
GDR	3890	65830	11.095	8.266
GHNA	71	5860	8.676	4.263
GMBA	0	90		
GNBS	0	70		
GNEA	21	750	6.620	3.045
GRCE	1430	21320	9.967	7.265
GRND		40		
GTML	44	3590	8.186	3.784
GYNA	11	400	5.991	2.398

HATI	9	850	6.745	2.197
HGKG		7700		
HNDS	18	1050	6.957	2.890
HNGR	1420	22690	10.030	7.258
ICLD	0	1320		
INDA	3310	85960	11.362	8.105
INDS	1050	29120	10.279	6.957
IRAN	7760	55510	10.924	8.957
IRAQ	1850	13880	9.538	7.523
IRLD	102	7470	8.919	4.625
ISRL	4160	13160	9.485	8.333
ITLY	4440	156590	11.961	8.398
IVCT	53	3630	8.197	3.970
JMCA	16	2270	7.728	2.773
JPAN	4780	496260	13.115	8.472
JRDN	144	1240	7.123	4.970
KMPC	68			
KNYA	51	2970	7.996	3.932
KORN	729	7100	8.868	6.592
KORS	991	19850	9.896	6.899
KWAT	235	15270	9.634	5.460
LAOS	19	300	5.704	2.944
LBNN	136	3290	8.099	4.913
LBRA	4	640	6.461	1.386

LBYA	201	13510	9.511	5.303
LSTO	0	190		
LXBG	23	2150	7.673	3.135
MALI	14	530	6.273	2.639
MDGS	28	1720	7.450	3.332
MLDV		10		
MLTA	2	460	6.131	0.693
MLWI	5	660	6.492	1.609
MLYS	515	9340	9.142	6.244
MNGL	74	1250	7.131	4.304
MRCO	253	7860	8.970	5.533
MRTN	8	420	6.040	2.079
MRTS	1	540	6.292	0.000
MXCO	528	63200	11.054	6.269
MZBQ	0	1640		
NCRG	32	1580	7.365	3.466
NGER	4	590	6.380	1.386
NGRA	1070	25600	10.150	6.975
NPAL	9	1340	7.200	2.197
NRWY	847	27110	10.208	6.742
NTHL	2660	78550	11.271	7.886
NZLD	262	13130	9.483	5.568
OMAN	655	1790	7.490	6.485
PERU	621	11670	9.365	6.431

PHLP	402	15930	9.676	5.996
PKST	622	11270	9.330	6.433
PLND	5090	88320	11.389	8.535
PNMA	15	2150	7.673	2.708
PPNG		1290		
PRGY	23	1470	7.293	3.135
PRTG	1000	15060	9.620	6.908
PRTR		7120		
QTAR	106	2200	7.696	4.663
RMNA	2230	26450	10.183	7.710
RWND	7	430	6.064	1.946
SAFR	1520	32270	10.382	7.326
SDAN	121	4140	8.328	4.796
SDAR	1750	33240	10.412	7.467
SMLA	21	340	5.829	3.045
SNGL	29	1800	7.496	3.367
SNGP	305	5510	8.614	5.720
SPAN	2820	97140	11.484	7.944
SRLE	5	610	6.413	1.609
SRLK	23	2540	7.840	3.135
SRNM	0	500		
STPR		40		
SWAZ	0	220		
SWDN	1980	66830	11.110	7.591

SWTZ	964	53840	10.894	6.871
SYCH		30		
SYRA	837	5330	8.581	6.730
TLND	398	14600	9.589	5.986
TNSA	65	4090	8.316	4.174
TNZN	65	2440	7.800	4.174
TOGO	8	560	6.328	2.079
TRKY	1600	36030	10.492	7.378
TRNT	6	2170	7.682	1.792
TWAN	1410	14890	9.608	7.251
UAR	59	8880	9.092	4.078
UGND	78	2680	7.894	4.357
UK	10200	211700	12.263	9.230
UPVL	13	640	6.461	2.565
URGY	73	3600	8.189	4.290
USA	91000	1519890	14.234	11.419
USSR	119000	649470	13.384	11.687
VNM				
VNMN	310			
VNMS	465			
VNZL	539	27320	10.215	6.290
WSMA		50		
YGSL	1600	33080	10.407	7.378
YMNA	37	410	6.016	3.611

YMNS	52	1210	7.098	3.951
ZAIR	143	3450	8.146	4.963
ZIMB	81	3460	8.149	4.394
ZMBA	94	2090	7.645	4.543

What is the correspondence between the quantity of labor (number of employees) and the quantity of capital (assets) among petroleum refiining companies?

From: 1996 FORTUNE 500.					
Copyright 1996 Time, Inc.					
All Rights Reserved.					
Fortune is a registered mark of Time, Inc.					
COMPANY	REVENUES	PROFITS	ASSETS	EMPLOYEES	INDUSTRY
Name	$ millions	$ millions	$ millions		
Exxon	110,009.0	6,470.0	91,296.0	82,000	Petroleum refining
Mobil	66,724.0	2,376.0	42,138.0	50,400	Petroleum refining
Texaco	36,787.0	607.0	24,937.0	28,247	Petroleum refining
Chevron	32,094.0	930.0	34,330.0	43,019	Petroleum refining
Amoco	27,665.0	1,862.0	29,845.0	42,689	Petroleum refining
USX	18,214.0	214.0	16,743.0	42,774	Petroleum refining
Atlantic Richfield	16,739.0	1,376.0	23,999.0	22,000	Petroleum refining
Phillips Petroleum	13,521.0	469.0	11,978.0	17,400	Petroleum refining
Ashland	11,251.1	23.9	6,991.6	32,800	Petroleum refining
Coastal	10,223.4	270.4	10,658.8	15,500	Petroleum refining

Sun	8,370.0	140.0	5,184.0	11,995	Petroleum refining
Unocal	7,527.0	260.3	9,891.0	12,509	Petroleum refining
Amerada Hess	7,524.8	(394.4)	7,756.4	9,574	Petroleum refining
Tosco	7,284.1	77.1	2,003.2	3,750	Petroleum refining
MAPCO	3,310.0	74.7	2,293.3	6,204	Petroleum refining
Valero Energy	3,019.8	59.8	2,876.7	1,658	Petroleum refining
Diamond Shamrock	2,956.7	47.3	2,245.4	11,250	Petroleum refining
Kerr-McGee	2,928.0	(31.2)	3,232.0	3,976	Petroleum refining
Ultramar	2,714.4	69.6	1,971.3	2,800	Petroleum refining
Pennzoil	2,490.0	(305.1)	4,307.8	9,758	Petroleum refining

“*r*”:

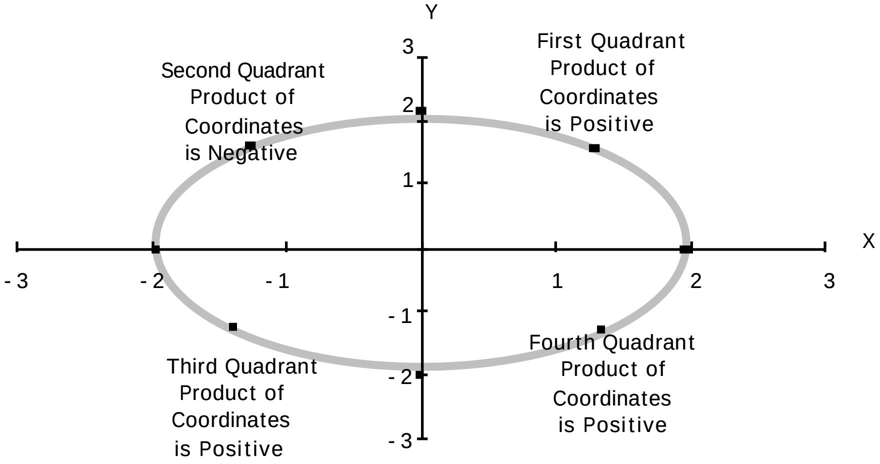
The Measure of Correlation

When a data analysis has been accomplished, and when the result is worth communicating, it becomes necessary to write a report. Reports are not as detailed as the work. And your reader may not be particularly fascinated by the details of the real work of hypothesis construction, examination of residuals for their pattern, revisions of hypotheses, and so forth that led to the result that merits a report. Creativity is a wonderful and peculiar process, but ultimately work will be judged by the result not than the process. At this point convention has greater value than it does in the creative process itself. And, where it is appropriate this is the time to use a conventional measure of the strength of the correlation between two variables. It is not as useful, as powerful, or as subtle as the examination of residuals, but it is conventional to summarize linear correlation between two variables with a number “r”.

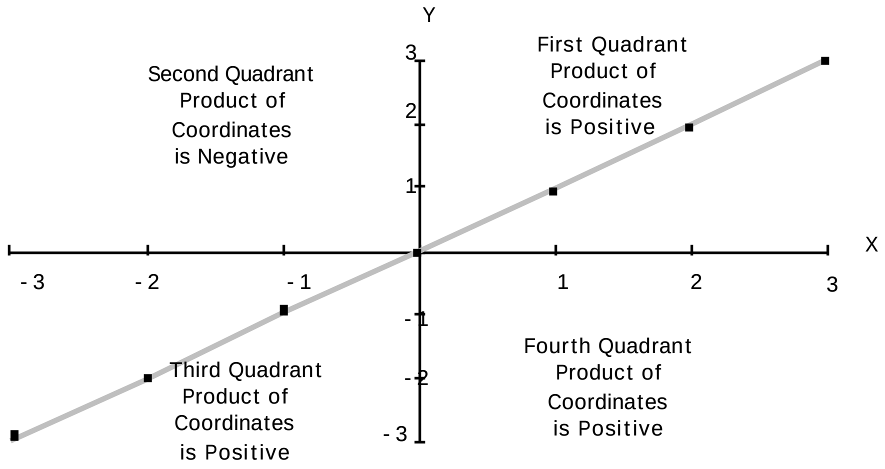
“r” is something of a magic number in statistics because r shows up in at least three contexts, three contexts where the same number makes sense for three different purposes. Here, I will introduce r as a measure of correlation, taking the pedagogical path of introducing r as if for the first time — as an answer to the question “How strong is the correlation?”

Just about the only kind of correlation that statistical technique is well prepared to talk about is a straight line correlation. And there are three things to be said about any straight line. In statistics, as in geometry, a line has an intercept and it has a slope. What's left, in data analysis, is the strength of the correlation or, to use one of the conventional terms, the "goodness of fit". A single number representing correlation is a clumsy weapon as compared to the human eyeball inspecting a pattern of residuals, but it is conventional, and no conventional journal article that uses linear correlation will be without conventional numbers. (With luck, you can use the residuals, and show the residuals, especially if there is a pattern — the serious reader will appreciate the fact that you have to satisfy two audiences, the reader who looks for convention as well as the serious reader.)

The basic intuition is that this collection of data points

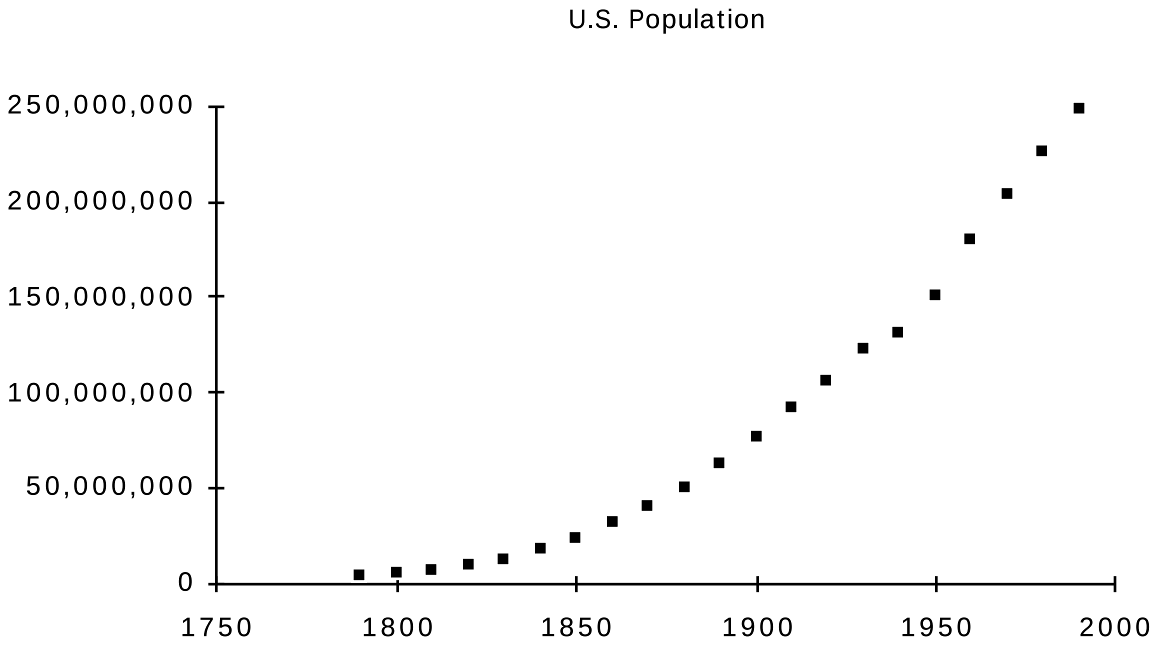


is not a line. While this collection of data points

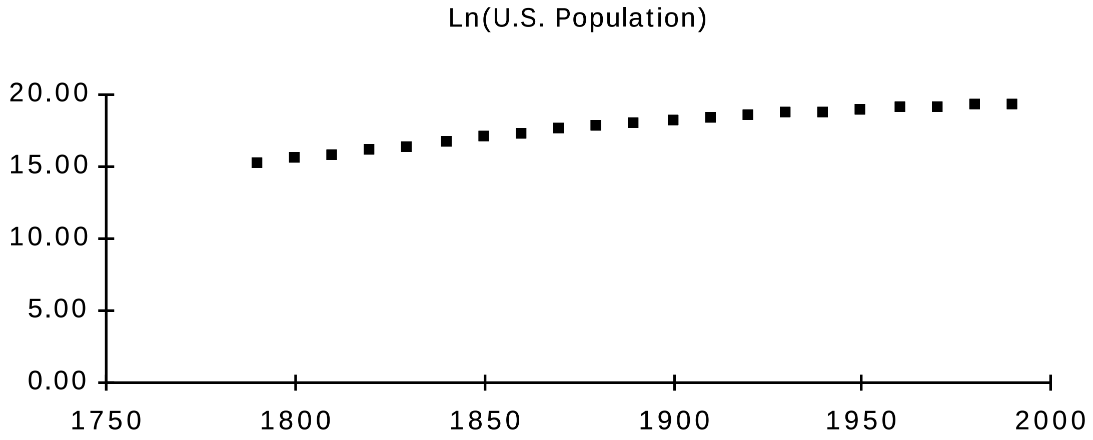


is a line (shown with six data points and a line. The data points are supposed to represent reality, while the line is supposed to represent what our fertile human imaginations would to "see" in those data).

That is what I am thinking about as some sort of ideal of linear data. But this is the kind of stuff we actually look at:



The first thing you do with such data is pay some attention to their behavior. A reasonable guess for a transformation of population to a well behaved variables is the logarithm, getting us something initially more linear looking, like



Now, after you've done all the real work of hypothesis construction, examining the residuals for pattern, revising hypotheses, and so forth, it comes time to write a report. Reports are not nearly as detailed as the work and, for the report, it is useful to put forth a summary statistic that answers the question "How linear?"

To invent a number that will answer that question, "How strong is the linear correlation?", I think, "What property of Figure __ (the simple line) can I summarize in a number. When I look at something truly linear, the number should say "good"; when I look at something truly non-linear (such as a circle) the computed number should say bad (bad meaning -- not at all like the number I get for a line).

The property I am going to work with, the basic intuition, is that a positively-sloped line will have lots of data points in quadrants I and III, while it will have very few data points in Quadrants II and IV. Fortunately, quadrants I and III have a common property that allows me to detect dominance of these two quadrants. Specifically, the product of two coordinates in these quadrants is positive while, in the opposite quadrants, in II and IV, the product of two coordinates is negative. So I can invent a number that adds up the dominance of I and

III as compared to II and IV by adding up the products of coordinates. . This is not yet the right answer, not yet, but it is the right start.

Rough intuition for correlation: $\sum_{i=1}^{n} x_i y_i$

Standardizing for number of data points:

That is the basic intuition. But this would-be index of correlation has serious flaws that need cleaning up to make it useful. For example, data sets come in different sizes, they have different "n's". This index would misjudge them: Using this index twelve data points with exactly the same pattern as six data points would, nevertheless, get twice the value for correlation. So, the index is better if it is modified to present an average instead of a sum.

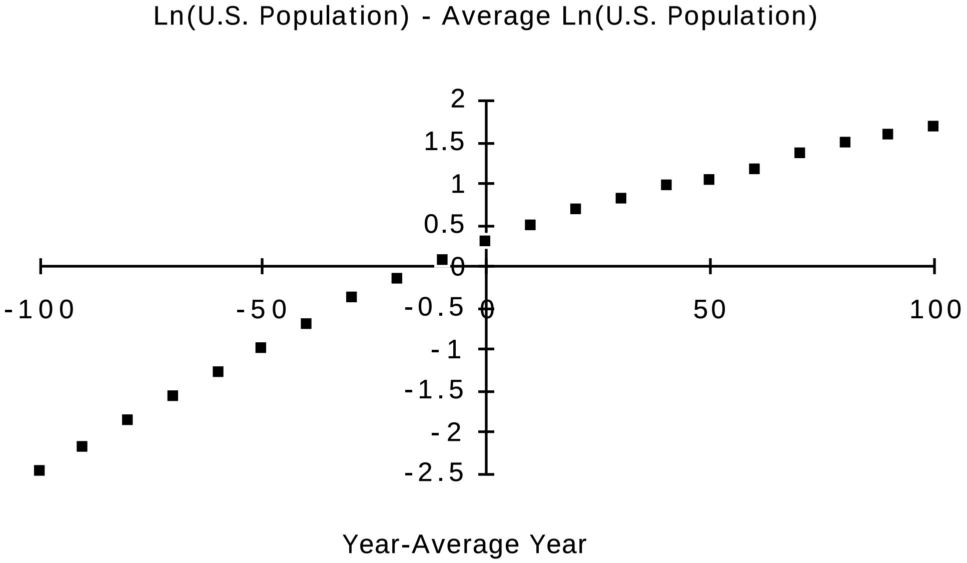
Rough intuition for correlation: $\frac{1}{n}\sum_{i=1}^{n} x_i y_i$

Standardizing for the "origin":

Another problem with this index is that data and even perfectly linear without ever passing through the origin. Like the population data above, the data could all lie in the first quadrant and still show correlation. Surely that's a problem — unless we care to conclude that all positive numbers are correlated. So the index of correlation can be improved by revising the data — moving the origin to the center of the data. Within the "least squares" framework, that center should be at the mean of x and the mean of y, leading to a modified index

Better intuition for goodness of fit: $\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$

Example with log population, translating the origin to the mean puts most of the data into proper quadrants (for the translated data).



Standardizing for Scale of the Variables:

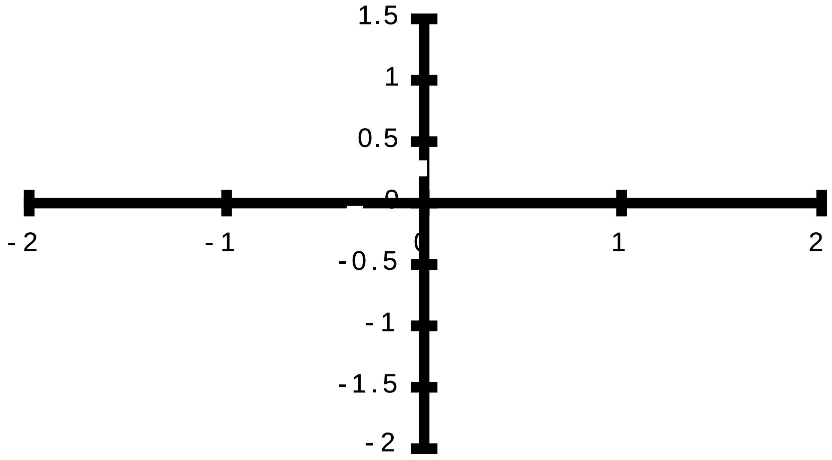
Finally there is a problem of scale. When you simply look at that graph you see a strong relation. But when you examine the numbers themselves, there is an enormous difference in scale between these two variables, +- 2.5 for one, +- 100 for the other. That doesn't affect the picture but it strongly affects this candidate for the index of correlation.

The index can be improved again by standardizing the scale: We standardize the "scale" of x by computing its standard deviation and then dividing by its standard deviation. The standard deviation for these numbers from 1790 to 1990 is 62. The standard deviation for these numbers from 15 to 19 is 1.309. But in standardized form, *subtracting the mean and dividing by the standard deviation*, both standardized variables have mean 0 and both standardized variables have standard deviation 1 — standardizing both variables to approximately the same scale. So, rewriting the numbers by subtracting the mean and dividing by the standard deviation I get

Better intuition for correlation: $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$

In graphical form for the population data:

(Ln Pop-Average Ln Pop)/Standard Deviation of Ln Pop



(fix graph, It is right, but make the point be making both scales have the same ticks and numbers on them -- Excel shrunk one scale ***)

There you see the relation (re-expressed) in what is called “standard form”. This example is typical of standard form graphs in that the graph is centered on zero, the ranges are much the same and, in fact, they will usually run in a range of plus or minus two standard deviations of the mean. The "imperfection" of this particular correlation (which is not linear) has expressed itself by a little bit of "leakage" into the second and fourth quadrants — affecting the xy products (these will be negative) and numerically diminishing the overall measure of correlation.

Conventionally we call

$$X_i \equiv \left(\frac{x_i - \bar{x}}{s_x}\right) \text{ and } Y_i \equiv \left(\frac{y_i - \bar{y}}{s_y}\right)$$

standardized variables

With these standardizations, that's it: We use the average cross product of these standardized forms of the original variables as the measure of linear correlation, naming it "r":

$$r = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

This "r" is very very important in the standard least squares approach. Enjoy the simplicity of the concept and the simplicity of the equation (in standard form) — but remember the thread that got us through the data, in this case through years and log populations, to the simple "X" and "Y" and r. Remember the thread because when we get to the next step in the standard "least squares" approach you have to do two things: You have to use these standardized forms, and you also have to remember how to follow the thread back to the data. The number "r" is an abstraction. And, as always, the numbers must come back to the data.

Exercises

Most statistical programs will compute “r” for you. Don’t do that. For now, compute r by working through the detail: Compute the mean for each variable. Compute the standard deviation for each variable. Compute the standardized form for each variable. And compute the mean cross product for the pair of variables, that’s the correlation coefficient, r.

Consider the correlation between U.S. population and year. Compute the correlation coefficient, r, with and without the use of the logarithm of population. What do the correlation coefficients report?

Consider the correlation between body weight and brain weight. Compute the correlation coefficient, r, with and without the use of the logarithms of each variable. What do the correlation coefficients report?

r:
Regression

The term "regression" refers to one particular way of estimating a summary line to fit a cloud of data. Let me defer the origin of the term "regression" itself: There is a solid reason for the connotation of regressing or moving backward, but the immediate problem is to introduce this particular way of estimating a good line to represent a cloud of data. So far, here's what you know — and this part does not change: When you have a hypothesis stating that one variable is a linear function of another variable, at least approximately, you generated expected values, the values that your "y" variable would have if the linear hypothesis were correct. Then you examine the exceptions, the residuals, looking at both the size of the residuals and their pattern. You use the size of the residuals both to keep score, how well does the linear hypothesis fit the data, and as a practical device to be used for finding the best slope and the best intercept.

The practical operation for finding the best fit is tedious in the extreme. I've used this tedious procedure for two reasons. Primarily, it absolutely forces you to look at the data. And that, in turn, leads to hypothesis construction, to treating a first data point or a last data point as an exception not related to the linear hypothesis, to breaking the curve into pieces, … to all sorts of intelligent but customized approaches responsive to the problem at hand. Looking at the data, very closely, instead of just committing the data to the computer, programming a set of predetermined questions, and writing up the results, simply leads to better data analysis.

The other reason for the tedious procedure is so that now, when I am prepared to drop the tedium, you will know, nevertheless, exactly what is going on. In most of the examples the procedure was to estimate an intercept and a slope and then compute the mean squared residual determined by this intercept and slope. Then I re-estimated the intercept and re-estimated the slope and re-examined the mean squared residual, accepting new estimates of the slope where the new estimates reduced this mean squared residual. That procedure *is* “regression” analysis in everything except name. However it benefits greatly for a straightforward application of the calculus. One thing that the calculus can be very good at is “optimizing”. Instead of the correct but tedious procedure you have been using, we can try to set up a measure of error which is suitable grist for the optimizing devices of the calculus and say, to our calculus, “optimize: find the particular slope and the particular intercept for which the size of the typical residual is reduced to a minimum.” This does not treat the important question of the pattern of the residuals — that remains the proper work for the human eye, and that is why you must look at the graphs even where the calculus can be relied upon to remove the tedium.

The calculus can not optimize just any measure of the residuals but it is excellent at optimizing with respect to the mean squared residual, working in the context of least squares. Here we call upon the calculus to solve a two variable problem, finding a line at the center of a cloud of data. And the procedure is exactly analogous to what we did earlier, finding the center for the values of one variable. Recall that for one variable, the “center” is a point to which the data are close. Using “close” in the sense of least squares implied that the mean was the center of the distribution. Using “close” in the sense of minimum absolute deviation implied that the median was the center of the distribution.

, using close in that is close to the stuff of which it is a center. “Close” can mean different things and the meaning that is historically easiest to work with — when computers are non existent and calculus is well developed — is “close in the sense of least squares”.

More formally, for one variable the average size of the variation around the center (when it is minimized) is called the variance of x:

So the measure of the center that is close to the data in the sense of least squares is the average.

$$V[\$x\ ] = \frac{1}{n}\sum_{i=1}^{n}\left(\$\text{x} - \$\,\overline{\text{x}}\right)^{2}$$

and the center, defined as the point of minimum variance was the mean.

$$\$\overline{x} = \frac{1}{n}\sum_{i=1}^{n}\$\text{x}$$

For interpretive work, when we need a number for spread around the mean that uses the same units as x, we use the standard deviation

$$s_{x} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\$\text{x} - \$\,\overline{\text{x}}\right)^{2}}$$

which is also called by the name “root mean squared error”, specifying the root, the average, and the squares that are visible in the formula.

I am going to use exactly the same logic to find a line which is the best line in the sense of least squares: I will define variation around the line and then choose the line with respect to which that variation is optimized. It is absolutely straightforward. But to make it look absolutely straightforward, I have to return to the use of standardized variables

$$X_{i} = \frac{x_{i} - \overline{x}}{s_{x}} \quad \text{and} \quad Y_{i} = \frac{y_{i} - \overline{y}}{s_{y}}$$

In terms of these standardized forms for x and y it was easy to construct the argument for measuring correlation as

$$r_{xy} = \frac{1}{n} \sum_{i=1}^{n} X_i Y_i$$

Continuing with these standardized forms of x and y it is straightforward to estimate the best linear summary of the relation between x and y. But remember the trail that leads back from the standardized variable to the original variables: If, for example I learn something about X (upper case "X", the standardized variable), then I have learned something about x (lower case "x", the original variable) with a little algebra to connect one to the other showing that

$$X_i = \frac{x_i - \overline{x}}{s_x} \Leftrightarrow x = s_x X + \overline{x}$$

which gets us back.

To find the best line, and to let the calculus eliminate the tedium, I construct the linear hypothesis

$$\widehat{Y}_i = M X_i + B$$

using a caret, "^" over the Y_i to indicate that this is not the true value of Y_i . It is the value that Y_i would have if it were truly predicted from the value of X_i. (with no residual). And now, just as I did for one variable, I create a measure of variation around the center (variation of the Y's around the line), and prepare to have the calculus find the best line.

$$E(M,B) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i \right)^2$$

The error variation depends on the choice of the line, which means that error is a function of the intercept B and the slope M. So, I need the value of B and the value of M for which E is smallest — the best fit line in the sense of least squares.

The details are a combination of algebra and basic calculus. Using the algebra, I want the M and B visible in the equation for error.

$$E(M,B) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \left[MX_i + B\right]\right)^2$$

And now I use the calculus: The minima of functions occur where their derivatives are equal to zero (or at an end point). So, noting that we are in a peculiar case where M and B are variables (while X and Y are constants -- they are whatever the data make them) — I will differentiate E with respect to B. I differentiate E with respect to M. I will set the two derivatives equal to zero and then solve these two equation as two simultaneous equations in two unknowns, M and B).

—— But first, let me establish a few things, “lemmas”, that will make the calculation simple: First I need to establish the average of a standardized variable. If that sounds peculiar to you, if it sounds peculiar to ask for the average of a thing for which I have no data, that’s good. It means you are thinking like a data analyst, which is what I wanted. But one of the mathematically pleasant (and useful) properties of standardized variables is that they have an average, always the same average, as a mathematical fact.

I could tell you the answer, but you should have the habit of proving these things for yourself, there is no need to look them up in the learned text of some expert. How do you get the answer to the question: What is the average of a standardized variable? You simply make the algebraic substitutions and simplify the result. So, starting with the definition of the average, here is the average, for any variable, standardized or not:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Now I make the algebraic substitution, using the definition of the standardized variable.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)$$

The rest is simplification: I "pull out the common factor" s_x

$$\overline{X} = \frac{1}{s_x}\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\right)$$

and then distribute the summation in order to add up the two terms separately

$$\overline{X} = \frac{1}{s_x}\left(\frac{1}{n}\sum_{i=1}^{n}x_i - \frac{1}{n}\sum_{i=1}^{n}\overline{x}\right)$$

And now, things are just about reduced: Inside the parentheses, the term on the left is, by definition, the mean of the original variable x .

$$\overline{X} = \frac{1}{s_x}\left(\overline{x} - \frac{1}{n}\sum_{i=1}^{n}\overline{x}\right)$$

Inside the parentheses, the expression on the right is adding a constant to itself n times, which means it is equal to

$$\overline{X} = \frac{1}{s_x}\left(\overline{x} - \frac{1}{n}n\overline{x}\right)$$

which simplifies to

$$\overline{X} = \frac{1}{s_x}\left(\overline{x} — \overline{x}\right)$$

which simplifies further to the answer

$$\overline{X} = 0$$

There is the result I need now in order to make subsequent computations simple. It tells me that any time I see $\overline{X}$ in an equation I can substitute the value 0.

That is peculiar if you are thinking about data: Here is a variable whose mean is always zero. But, of course, it is a standardized variable that was created expressly for the purpose of having a variable that was centered on its own average. All I've done with the mathematics is recover this fact about standardized variables by "proving" that the average (of a standardized variable) is 0.

In the same spirit, what is the variance of a standardized variable? If there is any doubt about the answer, I figure it out, as before, by making substitutions and simplifying the result. I begin with the definition. The variance of any variable, standardized or not, is

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Using what I have just established about the mean of X, I use what I established to simplify the present equation.

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i\right)^2$$

Now I am ready for substitution, using the definition of X.

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)^2$$

This invites simplification by factoring-out the denominator within the parentheses

$$s_X^2 = \frac{1}{s_x^2}\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right]$$

And now if you will check your definitions, you will recognize the thing within the square brackets on the right. By definition, it is the variance. So:

$$s_X^2 = \frac{1}{s_x^2}\left[s_x^2\right]$$

From which it follows

$$s_X^2 = 1$$

There is the second result I need now in order to make subsequent computations simple. It tells me that any time I see s_X^2 in an equation I can substitute the value 1. (

With these two lemmas in hand, I am ready to work on the partial derivatives of E(M,B), to find the best fit line in the sense of least squared error with respect to the variable Y.

I will differentiate E with respect to B and E with respect to M, producing two expressions. Eventually I will set both expressions equal to 0 creating two simultaneous equations that I will solve for B and M. But first, I simplify. Substituting the formula for E, I write

$$\begin{cases} \dfrac{\partial}{\partial B} E(M,B) = \dfrac{\partial}{\partial B}\left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(Y_i - \left[MX_i + B\right]\right)^2\right) \\ \dfrac{\partial}{\partial M} E(M,B) = \dfrac{\partial}{\partial M}\left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(Y_i - \left[MX_i + B\right]\right)^2\right) \end{cases}$$

Knowing that the derivative of the sum is the sum of the derivatives, the pair of expressions becomes

$$\begin{cases} \dfrac{\partial}{\partial B} E(M,B) = \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\dfrac{\partial}{\partial B}\left(Y_i - \left[MX_i + B\right]\right)^2\right) \\ \dfrac{\partial}{\partial M} E(M,B) = \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\dfrac{\partial}{\partial M}\left(Y_i - \left[MX_i + B\right]\right)^2\right) \end{cases}$$

Using the chain rule to deal with the squares,

$$\begin{cases} \dfrac{\partial}{\partial B} E(M,B) = \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}2\left(Y_i - \left[MX_i + B\right]\right)\dfrac{\partial}{\partial B}\left(Y_i - \left[MX_i + B\right]\right)\right) \\ \dfrac{\partial}{\partial M} E(M,B) = \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}2\left(Y_i - \left[MX_i + B\right]\right)\dfrac{\partial}{\partial M}\left(Y_i - \left[MX_i + B\right]\right)\right) \end{cases}$$

Only one of the expressions inside the nested parentheses at the right has the variable B in it, and only one has an M, so the derivatives simple again, leaving

$$\begin{cases} \dfrac{\partial}{\partial B}E(M,B) = -2\left(\dfrac{1}{n}\sum_{i=1}^{n}\left(Y_i - \left[MX_i + B\right]\right)\right) \\ \dfrac{\partial}{\partial M}E(M,B) = -2\left(\dfrac{1}{n}\sum_{i=1}^{n}\left(Y_i - \left[MX_i + B\right]\right)\left(X_i\right)\right) \end{cases}$$

Eventually, I have to set both of these expressions equal to zero and solve the simultaneous equations for B and M. But first, I use the lemmas to simplify these expressions.

Recalling that the derivative of the sum (within the parentheses) is the sum of the derivatives And noting that differentiating with respect to B is particularly simple I get

$$\begin{cases} \dfrac{\partial}{\partial B}E(M,B) = -2\left(\left(\dfrac{1}{n}\sum_{i=1}^{n} Y_i\right) - \left(\dfrac{1}{n}\sum_{i=1}^{n} MX_i\right) - \left(\dfrac{1}{n}\sum_{i=1}^{n} B\right)\right) \\ \dfrac{\partial}{\partial M}E(M,B) = -2\left(\left(\dfrac{1}{n}\sum_{i=1}^{n} X_iY_i\right) - \left(\dfrac{1}{n}\sum_{i=1}^{n} MX_i^2\right) - \left(\dfrac{1}{n}\sum_{i=1}^{n} BX_i\right)\right) \end{cases}$$

factoring out some the M's and the B's

$$\begin{cases} \dfrac{\partial}{\partial B}E(M,B) = -2\left(\left(\dfrac{1}{n}\sum_{i=1}^{n} Y_i\right) - M\left(\dfrac{1}{n}\sum_{i=1}^{n} X_i\right) - B\left(\dfrac{1}{n}\sum_{i=1}^{n} 1\right)\right) \\ \dfrac{\partial}{\partial M}E(M,B) = -2\left(\left(\dfrac{1}{n}\sum_{i=1}^{n} X_iY_i\right) - M\left(\dfrac{1}{n}\sum_{i=1}^{n} X_i^2\right) - B\left(\dfrac{1}{n}\sum_{i=1}^{n} X_i\right)\right) \end{cases}$$

That shows recognizable terms which are means and variances of standardized variables

$$\begin{cases} \dfrac{\partial}{\partial B}E(M,B) = -2\left(\left(\overline{Y}\right) - M\left(\overline{X}\right) - B\left(\dfrac{1}{n}n\right)\right) \\ \dfrac{\partial}{\partial M}E(M,B) = -2\left(\left(\dfrac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) - M\left(s_X^2\right) - B\left(\overline{X}\right)\right) \end{cases}$$

That allows a sharp simplification to

$$\begin{cases} \dfrac{\partial}{\partial B}E(M,B) = 2B \\ \dfrac{\partial}{\partial M}E(M,B) = -2\left(\left(\dfrac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) - M\right) \end{cases}$$

And the one complicated looking term is also recognizable. The context is different, but this is the “r” that has been used, earlier, as a measure of correlation.

$$\begin{cases} \dfrac{\partial}{\partial B}E(M,B) = 2B \\ \dfrac{\partial}{\partial M}E(M,B) = -2(r - M) \end{cases}$$

Now I go back to the game plan: I set these two partial derivatives equal to zero.

$$\begin{cases} 2B = 0 \\ -2(r - M) = 0 \end{cases}$$

Now, finally, I am rarely going to find a pair of simultaneous equations more easily solved than these. The first tells me B=0. The second tells me M=r.

$$\begin{cases} B = 0 \\ M = r \end{cases}$$

(*** Figure out how to get the Equation editor to align those equations to the left.)

Don't miss the simplicity of these statements: They say that B is always 0. The intercept of the best line for standardized X and standardized Y is *always* 0 — as you saw it in the graph where the line passed at least close to zero.

And while the slope is not simple, it shouldn't be: Something in this equation ought to depend on data. It is the slope. And while it isn't simple it is very nice to see "r" recurring. Previously r was introduced as a measure of correlation. Here, the same r is the slope of the best fit line in the sense of least squares for predicting Y as a linear function of X:

The fruit of all the standardization and all the math is that, in this form, it is simple. In standardized form the best line (best in the sense of least squares) is always

Y = rX

And now, back to data: What does this statement about X and Y say about the data variables, x and y, with which I began this odyssey into the optimization of a straight line. The formula for the best line for x and y is more complicated, but I don't have to remember it. I just remember the standardized equation and make the substitutions. \

Starting with the basic equation:

$$\widehat{Y}_i = r X_i$$

and undoing the standardization

$$\frac{\hat{y}_i - \bar{y}}{s_y} = r\,\frac{x_i - \bar{x}}{s_x}$$

Or

$$\left(\hat{y}_i - \bar{y}\right) = r\,\frac{s_y}{s_x}\left(x_i - \bar{x}\right)$$

Cleaning it up, to show x multiplied by a slope and to show what is added (the intercept) at the end (enclosing the slope and the intercept in parentheses

$$\hat{y}_i = \left(r\,\frac{s_y}{s_x}\right)x_i + \left(\bar{y} - r\,\frac{s_y}{s_x}\bar{x}\right)$$

That gives you the best slope and the best intercept, best in the sense of least squares. Take your choice, tedious minimization on the spread sheet or one shot, here’s the answer, using the calculus to tell you how to convert the means, standard deviations, and the correlation coefficient r into the best answer.

Variance of Data
= Variance of Signal + Variance of Noise:

How good is the best?

We're not quite done: When I asked for the center of the data, center in the sense of least squares, I asked two questions. First I asked which center was the best. Then I asked, "How good?" and got the variance and the standard deviation as the answers. Here, first I asked what was the best line, best in the sense of least squares. Now I ask "How good?" And I will get the variance and standard deviation of the residuals (whose average is zero). In addition, there is a fringe benefit: The answer to that question is so "nice" in a mathematical sense that it leads to certain conventional presentations. It is not obvious that the mathematical "niceness" of the least squares method should way heavily among the priorities of the data analyst choosing a method but, in use, it does.

So, resuming the discussion. In order to choose a best fit line, best in the sense of least squares, here is what I minimized:

$$E(M,B) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$$

How big is it, at its minimum? To answer the question, I just substitute what I know

$$E(M,B) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \left[rX_i\right]\right)^2$$

and I simplify it — continuing to known attributes of standard variables (mean 0, variance 1).

Squaring

$$E(M\ ,B) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{\,2} - 2rX_iY_i + \left(rX_i\right)^2\right)$$

Distributing the summation through the terms

$$E(M\ ,B) = \frac{1}{n}\sum_{i=1}^{n}Y_i^{\,2} - 2r\frac{1}{n}\sum_{i=1}^{n}X_iY_i + r^2\frac{1}{n}\sum_{i=1}^{n}X_i^{\,2}$$

And in this form I see the now-familiar terms. The term on the left uses the variance of a standardized variable. That's one. The term on the right uses the variance of a standardized variable. That's one. And the term in the center uses the mean product, for which we have the symbol r. So

$$E(M\ ,B) = 1 - 2rr + r^2$$

And that is the size of the error: It is always

$$E(M\ ,B) = 1 - r^2$$

Once again, the story is told by r. r is the measure of correlation. r is the slope of the regression line in standard form. And, now, r is the key quantity in assessing the size of the error.

The fringe benefits of this equation lead to mathematical nice results. For example, from this equation for the size of the error, it follows that r has absolute limits of minus one and plus one. This follows because I know one thing for sure about the *squared* error: Squared error can not be negative. Therefore zero is less than or equal to the error

$$0 \le 1 - r^2$$

And that equation tells me about the limits of correlation: Because zero is less than or equal to one minus r-squared,

$$r^2 \le 1$$

And as a consequence r itself is bounded by the interval plus-or-minus 1

$$-1 \le r \le 1$$

I can also figure out the limits of a bad correlation. How weak can a correlation be? You would like the English, “no correlation” to correspond to a mathematics that says “zero correlation”. But if it is true, it has to be proved. So, suppose that X doesn’t help at all as a predictor of Y? Suppose that I always predict $\overline{Y}$ regardless of X. Suppose that whatever the value of X, I always predict that Y will be equal to its mean — the number that is close to all the y’s, but gets no help from x? In this sad case the error is

$$E(M ,B) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$$

This formula is recognizable: Assuming that the prediction is always $\overline{Y}$, this formula for the error is identical to the variance of the standardized variable Y. The error is . And since the error is 1, I substitute this value into the equation and it determines a value for r. Thus, in the case that X is useless for predicting Y,

$$1 = 1 - r^2$$

And

$r = 0$

That gives me the value of r for no correlation. I can also ask for the value of r when the line is a perfect fit to the data (no non-zero residuals). The value of r had better be plus or minus 1, but again I have to prove it. So, suppose that the error vanishes

$$0 = 1 - r^2$$

Sure enough, solving that equation for r, r must be plus or minus 1.

And finally, one more piece of the puzzle that leads to the standard jargon: Finally, I want to look at the three pieces: Data, Signal, and Noise and show that the pieces add up. I already know

Data = Signal + Noise

That was an assumption. Now I am going to demonstrate something that is not an assumption. I am going to demonstrate that the *variance* of the data is equal to the *variance* of the signal plus the *variance* of the noise. When I can demonstrate that I will be able to phrase sentences that sound very good to the data analyst. I will be able to say how much of the data is explained by the hypothesis about the signal — which means "How much of the variance of the data is the variance of the hypothetical signal?" or letting the language drift into conventional form "What percent of the variance is *explained* by the hypothesis?"

The first step is to demonstrate the equation

Variance of the Data = Variance of the Signal + Variance of the Noise

(where "variance of the signal" means variance of the values that would be predicted using a linear equation for y as a function of x).

The variance of the data is simply the variance of Y, the variable we are trying to describe as a function of X. So

$$\text{Variance of Data} = \frac{1}{n}\sum_{i=1}^{n} Y_i^{2}$$

Now let me add and subtract a useful term that leaves the sum unchanged:

$$\mathit{Variance\ of\ Data} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - rX_i + rX_i\right)^{2}$$

Regrouping

$$\mathit{Variance\ of\ Data} = \frac{1}{n}\sum_{i=1}^{n}\left(\left[Y_i - rX_i\right] + rX_i\right)^{2}$$

and then squaring, I get three pieces

$$\mathit{Variance\ of\ Data} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - rX_i\right)^{2} + 2r\frac{1}{n}\sum_{i=1}^{n}X\left(Y_i - rX_i\right) + \frac{1}{n}\sum_{i=1}^{n}\left(rX_i\right)^{2}$$

Now two of these three pieces look familiar: The term on the left is the variance of the residuals. The term on the right, with components rX_i , is the variance of the predicted values of Y (around an average prediction of zero). That leaves "stuff" in the middle which, we hope, is zero.

Variance of Data = Variance of Error + stuff + Variance of Signal

But it must be checked. So, starting at 1/n

$$\frac{1}{n}\sum_{i=1}^{n} X_i\left(Y_i - rX_i\right) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\sum_{i=1}^{n} rX_i^2$$

I recognize both of these expressions, factoring r from the second term

$$\frac{1}{n}\sum_{i=1}^{n} X_i\left(Y_i - rX_i\right) = r - r\frac{1}{n}\sum_{i=1}^{n} X_i^2$$

from which it follows

$$\frac{1}{n}\sum_{i=1}^{n} X_i\left(Y_i - rX_i\right) = r - r$$

and

$$\frac{1}{n}\sum_{i=1}^{n} X_i\left(Y_i - rX_i\right) = 0$$

This gets rid of the middle term, its value is zero, and leaves the target equation

Variance of Data = Variance of Error + Variance of Signal

or

$1 = (1\text{-}r^2) + r^2$

where

the Variance of Data is 1

the Variance of Error is $(1\text{-}r^2)$

and

the Variance of Signal is r^2

or in another technical expression for the same thing

Total Variance = Unexplained Variance + Explained Variance[1]

This leads to phrases that you will read over and over again in statistical reports, phrases like “The correlation explains 40% of the variance.” The phrase refers to the fact that the total variance is 1 and that the unexplained variance and the explained variance add up to one. so if the variance of the predicted values is .40 and the variance of the residuals is .60, you may say (with somewhat dubious linguistic precision), 60 percent of the variation is unexplained. Or, the prediction explains 40 percent of the variance.

[1] Over and over again here I am on thin ice referring to these things as variances: After all, variances are variations around a mean, and I am not showing a mean in this formula. To clean it up, I have to show that the missing mean is zero.

Exercises

1. Look at the data for Brain Weight and Body Weight. Write a short statement describing the quality of the prediction in conventional terms, using r and $r^{2.}$

Now explore the limits of the conventional claims that one variable “explains” another:

2. Repeat the first problem using brain weight and body weight, without logs. Compare this to the first answer (using logs). The answers are different. Reconcile them.

3. Compute some hypothetical data where your x goes from 1 to 10 and your y increases multiplicatively (below). You know for certain that these y’s are not a linear function of these x’s — in this case y = $1.1^{(x-1)}$. But, as a thought experiment, be dumb: Use a linear equation, use regression, to predict y from x. Use r and r^2 to report how well x explains y. Reconcile the superficial implications of the r and r^2 with the fact that anyone claiming that this y is a linear function of this x has clearly failed to explain the relation at all. (The word “explain” is rich with ambiguity. It has many meanings.)

X	Y
1	1
2	1.1
3	1.21
4	1.331
5	1.4641
6	1.61051
7	1.771561
8	1.9487171
9	2.14358881
10	2.35794769

Choices:
The Line? *Which* Line?

The term “regression” refers to one particular way of estimating a summary line to fit a cloud of data. There are others. In fact, while persons beginning the study of data analysis sometimes think of it as an intellectual rock: “here is how it’s done”, the truth is that the methods themselves remain the subject of active research — particularly with regard to methods of matching summary lines to clouds of data. There is no way that a person can be a responsible data analyst without realizing that there are alternative methods on the menu and that there are choices to be made — responsibilities that can not be ducked by delegating the choice to the computer, or by adopting the choices of a previous writer, or by referring to a text book,. Choice should also teach the humility to avoid overstating what the data have “told” you: It is difficult to believe results to two or three decimal digits, and perhaps to base policy decisions on fine comparisons among results — if you know full well that different methods would have given somewhat different numbers and if you know full well that there is no clear and obvious argument proving that one is right and the other is wrong.

Among the options available for fitting lines to the data, the first and most obvious option is “none of the above”. Very few real world examples, allow a *routine* application of any line fitting technique. In earlier chapters the relation between fertilizer and crop yield (four data points), the relation between time and soy bean plant length (seven data points), and the relation between time and the size of the population of the United States (twenty-one data points) — were all routine prosaic examples. Yet none of the analyses would have been well served by dumping all of the data into the computer and waiting for regression procedures to come up with a description.

Another option is created by the choice between least squares statistics and minimum absolute deviation statistics. Minimum absolute deviation statistics are rarely used, as compared to least squares statistics, probably for reasons of custom and mathematical ease.

OLS versus OLS versus OLS :
Ordinary Least Squared Error for Y
Versus Ordinary Least Squared Error for X
Versus Orthogonal Least Squared Error

Another option is both more subtle and more drastic in its effect on data analysis. I used least squares to minimize the residuals of one of the variables, always “y”, always represented vertically on the graph. For the moment I acted as if that were the obvious and only thing to do. Now, let me demonstrate alternatives to the fitting of “y”.

To demonstrate let me use these stylized hypothetical data describing combinations of income and education for seven individuals.

Person	Years of Education	Income
1	12	$20,000
2	12	$30,000
3	16	$20,000
4	16	$30,000
5	16	$40,000
6	20	$30,000
7	20	$40,000

Alright: The one variable distribution is symmetrical in each case — likely I need no re-expression (Real data on income and education would not be so well-behaved — I am simplifying, as before, to make the point transparent.)

So, the means are _____; the standard deviations are _____. Then computing the standardized variables, and then computing the mean “cross product” of the standardized variables, I get r = .43.

Person	Years of Education	Income	Predicted Income	(Observed – Predicted)	Standardized Education	Standardized Income	Product of Standardized Variables
1	12	$20,000	$25,000	–$5,000	-1.3228757	-1.3228757	1.75
2	12	$30,000	$25,000	$5,000	-1.3228757	0	0
3	16	$20,000	$30,000	–$10,000	0	-1.3228757	0
4	16	$30,000	$30,000	$0	0	0	0
5	16	$40,000	$30,000	$10,000	0	1.3228757	0
6	20	$30,000	$35,000	–$5,000	1.3228757	0	0
7	20	$40,000	$35,000	$5,000	1.3228757	1.3228757	1.75
Average	16	$30,000					0.43
Stand dev.	3.024	$7,559.289					

Remembering that, in standardized form, the regression equation (best fit, predicting *Y* from *X*, best in the sense of least squares) is

$Y = r\,X$

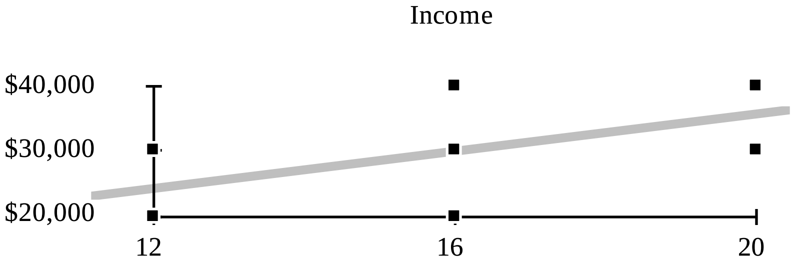
And then substituting to unstandardize:

$$y = \left(r\left(\frac{s_y}{s_x}\right)\right)x + \left(\bar{y} - \left(r\left(\frac{s_y}{s_x}\right)\right)\bar{x}\right)$$

So the regression equation for income as a function of education is:

Income = ($1,250 / year) * (Years of Education) + $10,000.

The regression equation says that a hypothetical person with zero years of formal education would have had an income of $10,000. Other people would have that $10,000 plus $1,250 for each year of formal education. The graph is as shown below.



But now I want to show you a problem: Not so much a problem as it is a requirement that you understand exactly what you are doing when you use a regression line. The analysis I have just completed tells you how much money you can expect (on the average) for another year in school. Answer $1,250.

But now, let me ask a different question of the same data. Now you are a consultant to an advertising agency. Your agency is working on different products, targeted to people with different levels of income. And so they ask you, if we are targeting consumers whose income is in the neighborhood of $20,000 range, what level of education should we expect of these consumers? And if, by contrast, we am targeting consumers with incomes in the neighborhood of $40,000 group how many years of formal education should we expect, on the average. Here it is

$X = r\ Y$

And then substituting to unstandardize:

$x = \left(r\left({}^{s_x}\!/_{s_y}\right)\right)y + \left(\bar{x} - \left(r\left({}^{s_x}\!/_{s_y}\right)\right)\bar{y}\right)$ So:

Years of Education = (.0002 years / dollar) * (Income) + 10.

Or, making a concession to my intuition, which has trouble with values like .0002 years per dollar, I'll write the equation in thousands of dollars:

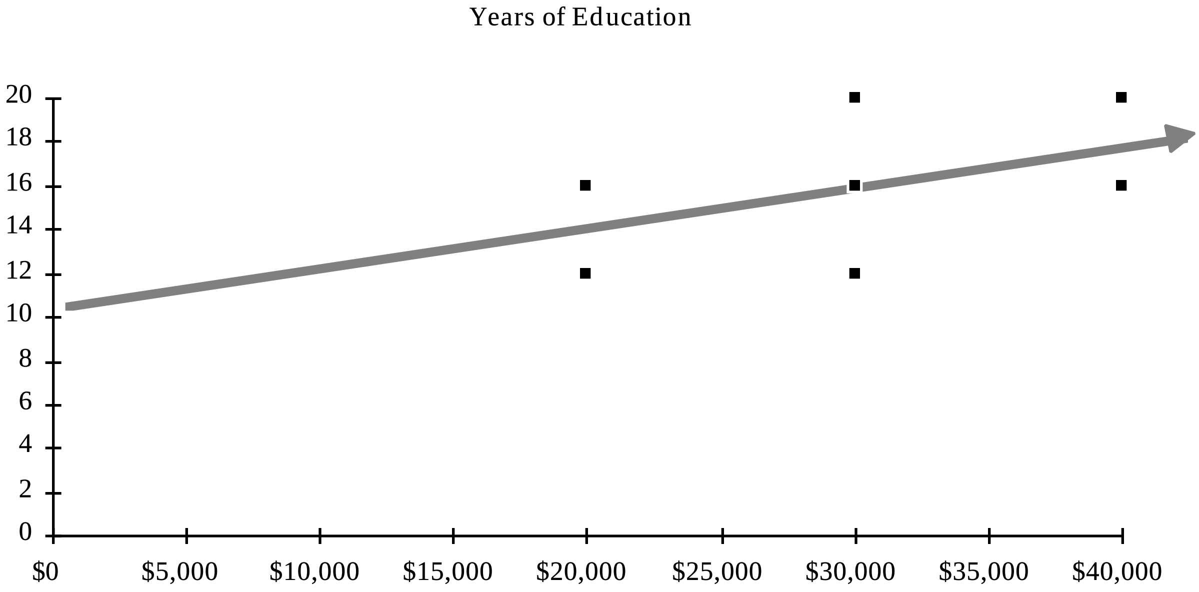
Years of education

= (.2 years of education per thousand dollars) * income (in thousands) + 10,85 years.

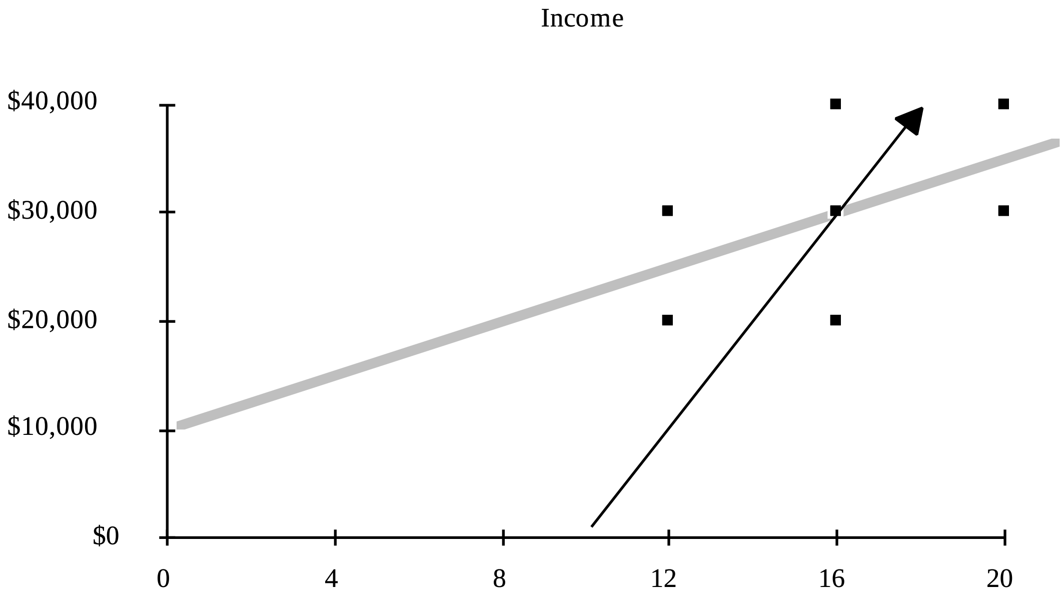
or in tens of thousands of dollars,

Years of education

= (2 years of education per ten thousand dollars) * income (in tens of thousands) + 10.85 years.



Now, I want to point out one not too minor detail: As you can see, I have switched axes: In one graph I have education left to right, used to predict income, bottom to top. In the last graph I have income left to right, used to predict education, bottom to top. Let me combine the two of these in one graph, superimposing one upon th other:



These are two different lines: Same data, different lines. When I posed two different questions of the one set of data I got two different answers.

That is very important because it means that neither of these regression lines is an objective description of the data. Neither regression line can claim to be "the facts and nothing but the facts — don't argue with the numbers." No, depending on the question that I, the analyst, chose to ask one of these answers, or the other of these answers (or neither of these answers) is correct.

Does it matter? Yes, a great deal. Expressing these two slopes in a common unit, one slope says $1,250 per year. The other says $5,000 per year of education — separated by a factor of four.

There is nothing wrong with either of these equations. Each is the right answer to the question that was asked. Discuss: If you want to predict Income from education, then you should minimize errors with respect to predictions of income. You will get the line with slope $1,250 per year. If you want to predict Education from Income, then you should minimize errors with respect to predictions of Education. You will get the line with slope 2 years per $10,000 (corresponds to $5,000 per year). But you have to be clear that prediction is not the same as objective description. (Again, neither is an objective description of the data.)

To make the point by reducing it to a possible absurdity, consider that we have data reporting the physical heights for pairs of brothers. Randomly, I will assign the height of one of the two brothers as “x” and assign the height of the other brother as y. Now, I ask you to use the height of one brother, which I will tell you, to predict the height of the other brother. Here is the pattern. If I tell you that one brother is 5’10”, about average for these hypothetical data, you can reasonable expect his brother to be about average, about 5’10”. If I tell you that one brother is 4’10” considerably below average, you can reasonably expect his brother to be considerably below average, but not so extreme. You might expect something like 5”. And if I tell you that one brother is 6’10”, considerably above average, you can reasonably expect his brother to be considerably above average, but not so extreme. You might expect something like 6’8”.

This is the logic of “regression”: If you have somebody who is above average, you expect a result that is above average — but a little closer to the mean: “regression toward the mean of the population”.

Now, let me try to confuse you while you try to resist: Suppose I report to you that one brother is 6’8”. You can reasonably expect this tall man to have a tall brother — but not quite so extreme: Perhaps 6’6”.

That means that men who are 6’10” would be predicted to have brothers who are, on the average 6’8”. And men who are 6’8” would be expected to have brother who are, on the average, 6’6”.

That sounds strange: It sounds like I am saying that a man who is 6’10” will have a brother who is 6’8”. But a man who is 6’8” will have a brother who is 6’6”. That is impossible, but that is not what I am saying: I am saying that the *average* in one case is 6’8” and the average in the other case is 6’6”. That is possible and that is regression analysis.

There is a third demand I can make of these data and it leads to a third answer. Suppose that these data are part of a larger data set, pairs of persons different by birth order, sex, nutrition, whatever. Then in this one case where the persons being compared are, in fact, the same except for random assignment to group “x” or group “y”, I want the data analysis to give me the numerical equivalent of the English statement that they are the same except for random variation.

The answer to this question come by thinking about “the line that is closest to the data”. This line is descriptive. It contains no built in assumption about predicting y from x or x from y. It is the line that is closest to the data. As in ordinary Euclidean geometry, the distance between a point and a line is the length of the perpendicular from the point to the line. This is the Orthogonal Least Squares Line and it has the equation

$Y = X$ (the r has vanished).

That looks too simple in standardized form, but restoring reality by unstandardizing, the descriptive (orthogonal least squares best line is)

$$y = \frac{s_y}{s_x} x + \left(\overline{y} - \frac{s_y}{s_x} \overline{x} \right)$$

When you do not, yourself, wish to impose order on the variables, when you are not trying to predict one from the other. choose the orthogonal least squares line to treat the variables symmetrically. It is entirely possible to use several procedures on the same data — you protect yourself and your reader by saying what you have done and providing sufficient data and numbers to allow the reader both to reproduce what you have done and to complete alternative analysis. This is not simply a matter of courtesy. It is a necessity because in perfectly ordinary data different estimates of slope, for example, will differ from each other by factors of three or ten or more — depending on the choice of method. You maintain integrity, first, by making it absolutely clear what you have done and, second, by equipping your reader to explore the paths you have not taken.

Education and Income 1993: Stream of Evidence

One of the truisms of modern life is that higher education is associated with higher economic return to the individual. It is, of course, a complicated relation: Even the averages will be affected by age, sex, occupation, parents' backgrounds and other variables. And within each group defined by age, sex, occupation, and so forth, there will be considerable variation around the average. But the general nature of the relation, more education more money, is so widely understood and, presumably, so strong that it is worth beginning with the unadorned data: Education by Income. These data are from the 1993 General Social Survey from the National Opinion Research Center. The data for 1607 adults indicate education and income (plus about 400 other indicators), where "education" is defined as years of school completed, and the income indicator used here is the respondent's personal income (not household income or wealth). The responses for education range from zero years of education to 20 while income is presented in twenty-one categories ranging from zero income to $75,000 per year-plus.

Here are the first 10 rows of my spread sheet. It demonstrates some of the idiosyncrasies of the culture of data analysis. (Data analysts, like another other profession have a culture, and like any culture it develops lags, which are holdovers from older technologies and just plain un-thought-out practices that have become customary.)

Note, for example, the names of the variables "EDUC" and "RINCOM91". Not very long ago the costs of most things involved with computers were so high that it was the practice to cut corners. Little things, like exclusive use of upper case letters, saved money. Conventionally names were cut to a maximum of 8 characters, hence "EDUC", rather than the English "Education", and "RINCOM91", which is the Respondent's Income in 1991. Add to that the advantage of giving variables the same name this year as you they were given last year and the year before, and the result is the chopped English commonly used for the names of variables.

Item	EDUC	RINCOM91	Lookup Income	Ln of Income (base e)
1	16	18	$45,000	10.71
2	12	14	$23,750	10.08
3	12	16	$32,500	10.39
4	14	21	$100,000	11.51
5	14	21	$100,000	11.51
6	15	17	$37,500	10.53
7	15	13	$21,250	9.96
8	12	13	$21,250	9.96
9	17	3	$3,500	8.16
10	12	15	$27,500	10.22

For the same reasons of practical necessity, necessity in an earlier era of computing, you didn't write $45,000. That would take seven characters, including the comma. Instead, you wrote "18". The meaning of "18" offers nothing to the intuition of the data analyst, but "18" as compared to "$45,000" saves five characters. So it was the practice to write "18" in the data set and create a look up table (in a "codebook") that decoded the symbols into data. Here for example, the lookup table was:

0 M NAP
1 LT $1000
2 $1000-2999
3 $3000-3999
4 $4000-4999
5 $5000-5999
6 $6000-6999
7 $7000-7999
8 $8000-9999
9 $10000-12499
10 $12500-14999
11 $15000-17499
12 $17500-19999

13 $20000-22499
14 $22500-24999
15 $25000-29999
16 $30000-34999
17 $35000-39999
18 $40000-49999
19 $50000-59999
20 $60000-74999
21 $75000+
22 REFUSED
98 M DK
99 M NA

Note the intentional lack of correspondence between the symbol "0" and an income of zero or between the symbol "0" and the lowest category of income. This custom dates back to a time when the absence of data might leave no mark on the data sheet. It is absolutely routine for respondents to refuse one or more questions (and, for that matter, to refuse the whole questionnaire). So you have to be prepared to distinguish between no datum and no income and you have to be prepared to catch mistakes when the distinction breaks down. So, it seemed wise to minimize errors by not using 0's as data. Combine this confusion with an equally careful practice of intentionally attaching nonsense symbols to nonsense data and you have a mess. (Introducing "99" for No Answer means that if, somehow, that symbol "99" were used as data it would stand out in a stem and leaf or in a two-variable graph. The data analyst will see the steam and leaf, or the graph, know that something is very wrong, and quickly track down the error.)

There is no obvious convention for these symbols, so each study requires you to "look up" the symbols in the codebook and find out what they stand for. Here the symbol "0" stands for "M NAP", that is, "Missing - Not Applicable." Here "22" stands for "Refused", "98" stands for "Missing - Don't Know", and "99" stands for "Missing - No Answer."

You want all of this detail about "non-responses" for three reasons. First, you want to know how many people responded. The question may have been asked of 1,500 people. That does not mean that 1,500 people responded. Second, non-responses are very likely be associated with particular values (usually extreme values) which would mean, in this case, that very low incomes and very high incomes are less likely to be recorded). And third, the reasons for non response are likely be associated with other variables. So you have to be alert to the possible divergence between *observed* correlations (using the data) and true but unknown correlations that you *would* have seen if everyone had responded.

You also see the custom of placing numbers like income in categories. There are arguments for and against the use of categories rather than raw income numbers. But whatever the argument, the use of categories, rather than income numbers, has consequences. In this case it means that all people with incomes above or equal to $75,000 dollars are the same for purposes of analysis: $75,000, $100,000, $1,000,000 — all equal. We can argue whether or not that is an appropriate decision (a debate sometimes described as the "One person, one vote" versus the "One dollar, one vote" debate), but with these data the decision has been made, the detail is gone, and it can not be restored.

Now, on to the analysis or, at least to the preparations for the analysis. Using my spread sheet program's "lookup" function I have re-expressed income with approximate values appropriate to each category. I have substituted a nonsensical -999999 for all forms of missing data, and then removed all missing data to the bottom of the spread sheet by sorting the data. For income that removes 539 of the 1607 respondents, about one-third missing. For education that removes 8 respondents.

These deletions are certainly disconcerting and hard to check without analyzing the attributes of these people with respect to other variables (Are they disproportionately men or women? Are they disproportionately young, old, or in some other age group? Are they simply unemployed or unemployed outside the household?) The briefest check, within these data shows that the 539 missing income population has a mean education of 12.0 years of education as compared to 13.1 years for the whole 1607 population. Their standard deviation for education is 3.1 as compared to 3.0. That tells me that the histograms of the two distributions could look very similar, with the distribution of incomes offset to lower education. And although the difference is small, it is probably real, not the luck of the draw.[1]

[1] I find that I am resorting to statistical knowledge that I have not included in the text. So, let me explain, or at least start to explain the difference between the standard deviation of the values of education and the standard deviation of *means* of the values of education. The mean education in this population is 13.1 years. That is a fact resulting from the obvious computation. The standard deviation of values around this mean is 3.0 years. Again, this is a fact resulting from the obvious computation.

For analysis of the data using both education and income, this reduces the number of adults from 1607 to 1017, removing 40% of the data. Presuming that these "missing" respondents are not a random sample of all respondents I am worried. But I have no effective way of

---

But consider: Suppose that I went out and collected new data for another 1607 people. I would get another mean and it would almost certainly be close to but slightly different from the mean I found in the first sample of 1607. If I did this again and again, getting a new sample and computing the mean in each sample I would get a lot of means, most of them close to the original but slightly different.

These means themselves would have a mean and a standard deviation. And if I were comparing one sample to another, asking whether one of these samples has a mean that is too much larger than the others, or too much smaller than the others to be ignored, then I would also need to know the standard deviation of the mean.

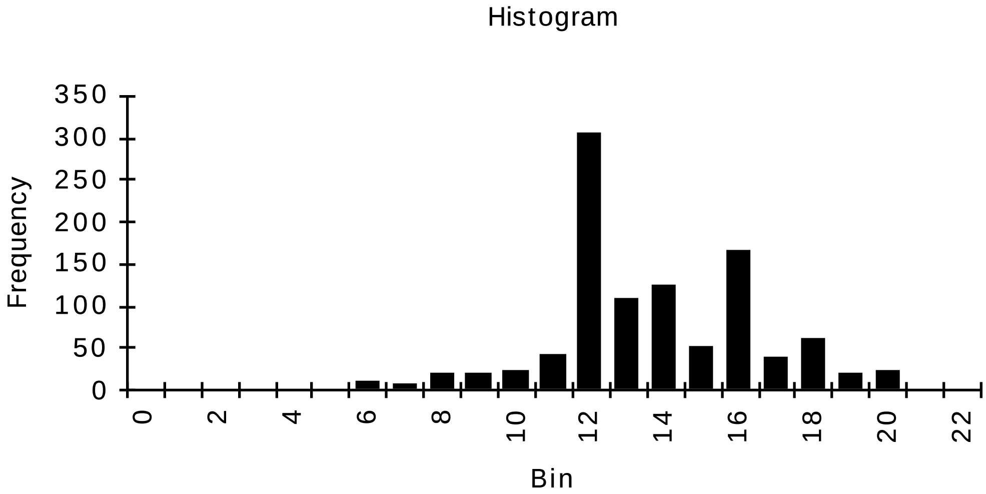
That is close to what we are doing here. We are comparing one group of 1607 with a mean of 13.1 years of education to a subset of 539 people who have a mean of 12.0 years of education. Are these two means close? I narrow the question by asking whether or not these two means are close as compared to the standard deviation *of the mean*.

Fortunately, statistics is able to estimate the standard deviation of the mean without the need to actually perform the experiment, without actually collecting new data and computing the mean, again and again and again. We know that the standard deviation of the mean is approximately the standard deviation of the values divided by the square root of the number of values. Here the standard deviation of the subset is 3.1. The number of values in the subset is 539. The square root of 1599 is 23.2. So the standard deviation of the mean is 3.1/23.2 = .13. That is my estimate for the standard deviation of the mean itself.

The difference between these two means is 1.1 year of education which is more than 8 times greater than the standard deviation of the mean. So, the difference in education between the subset and the whole is small, 1.1 year of education. But it is almost certainly a real difference, the difference is small but greater than I would expect just by the luck of the draw.

worrying about these numbers here. The correct way to check what is special about the missing people is to considering other variables available in the study — which is well beyond the scope of this exercise.)

The education variable (for the 1017 people) has decidedly non-bell-shaped exceptions:



These bumps are characteristic of education distributions, with bumps at the numbers of years that correspond to degrees. A bump at completion of grade school is no longer apparent in 1993. But there is a bump at 12, usually completing a high school degree in the U.S. and there is a bump again at 16, a college degree.

Is it symmetrical? That is a bit tricky because I don't really know how I want to qualify this question in order to accommodate these perfectly reasonable bumps. Let me take a look

Mid Values, Years of Education

Count, n=1015	From bottom	From Top	MidValue	
508	13	13	13	Median
254.5	12	16	14	Mid Quartile
127.5	12	17	14.5	Mid Eighth
64	10	18	14	Mid Sixteenth
32.5	7.5	19	13.25	Mid Thirty-Second

Mid Values, Square Root of Years of Education

Count, n=1015	From bottom	From Top	MidValue	
508	3.61	3.61	3.61	Median
254.5	3.46	4	3.73	Mid Quartile
127.5	3.46	4.12	3.79	Mid Eighth
64	3.16	4.24	3.57	Mid Sixteenth
32.5	2.73	4.36	3.54	Mid Thirty-Second

Mid Values, Natural Log of Years of Education

Count, n=1015	From bottom	From Top	MidValue	
508	2.56	2.56	2.56	Median
254.5	2.48	2.77	2.63	Mid Quartile
127.5	2.48	2.83	2.66	Mid Eighth
64	2.30	2.89	2.60	Mid Sixteenth
32.5	2.06	2.94	2.50	Mid Thirty-Second

There are enough values here to allow me to pursue quite a number of mid values without running out of data. The median, mid quartile and mid eighth do show a trend. But the mid sixteenth and mid thirty-second show the reverse.

I know that this is a bit of a mess, and I don't trust the top number — there are more than a few of us with greater than 20 years of education but we seem to have been left out or lumped in with our relatively uneducated friends with merely 20 years of education. So I think that the distribution has more of a tail than it is able to show with these categories. Checking to see whether a re-expression would fix up the trend among the first three mid values, the curious result is that all three sets of mid values show the same thing. So the test for symmetry is beautifully indeterminate, quite willing to accept the original numbers, or the square roots, or the logs, or more. That is no help at all. I will start simply, using years of education.

The same procedure applied to the income distribution of this limited population yields

Mid Values, Respondent's Income

Count, n=1015	From bottom	From Top	MidValue	
508	$21,250	$21,250	$21,250	Median
254.5	$11,250	$32,500	$21,875	Mid Quartile
127.5	$5,500	$45,000	$25,250	Mid Eighth
64	$2,000	$67,500	$34,750	Mid Sixteenth
32.5	$2,000	$100,000	$51,000	Mid Thirty-Second

Mid Values, Square Root of Respondent's Income

Count, n=1015	From bottom	From Top	MidValue	
508	145.77	145.77	145.77	Median
254.5	106.07	180.28	143.18	Mid Quartile
127.5	74.16	212.13	143.15	Mid Eighth
64	44.72	259.81	152.27	Mid Sixteenth
32.5	44.72	316.23	180.48	Mid Thirty-Second

Mid Values, Natural Log of Respondent's Income

Count, n=1015	From bottom	From Top	MidValue	
508	9.96	9.96	9.96	Median
254.5	9.33	10.39	9.86	Mid Quartile
127.5	8.61	10.71	9.66	Mid Eighth
64	8.16	11.12	9.74	Mid Sixteenth
32.5	7.60	11.51	9.56	Mid Thirty-Second

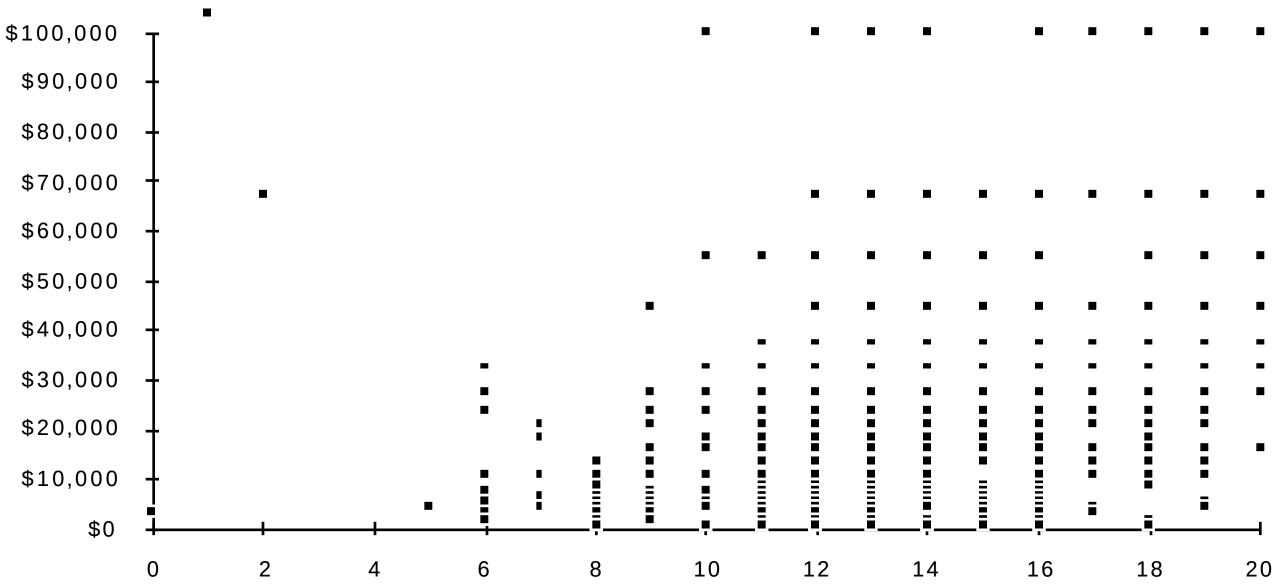
These numbers are strange: The distribution of incomes is not symmetrical. It has a tail toward the high values. That is reasonable. But the square roots of income lead to ambiguous indicators: The median, the mid quartile and the mid eighth drift slightly toward low values. The mid sixteenth and mid thirty-second move strongly toward the high values. Worse, the distribution of the logarithm of income indicates a tale to the left, if any. This is surprising: I have it in my head that income distributions will not be symmetrical measured in units of dollars but will be symmetrical measured in units of log dollars. That is what I expect. That is not what these data show. What's more, I am so sure of this that I question the data.

A close look at the data tells me that the unit of analysis in these data is the family, not the individual — it is a representative sample of family units from which the individual who is actually interviewed is chosen by ???? I don't know what. Either I am wrong about the shape of the U.S. income distribution or else there is something in these data that systematically under-represents high income individuals.

That tells me I am not going to reach a "final report" with these data: I will have to check both the income distributions and the educational distribution against other sources before I trust either distribution or their correlation. I'm also going to have to re-examine my own expectation about income distributions. When I think about income carefully, as I am now forced to do, I'm not sure what I should expect. After all, a very large part of the population has no cash income at all, and my intuition didn't cover that very realistic contingency. For the moment, if there is a bias I suspect that it is reducing the high income end of the distribution.

Proceeding rather tentatively, I am ready to look at the two variable distribution.

A graph of the relation between the two variables provides only a slight insight into the relation because of the grouping of the income data into categories. The grouping hides visual differences between the numbers of "dots" at each point on the graph. But the graph suffices to bring home the fact that there is a wide range of incomes across the entire range of educational achievement, particularly among those who have completed high school, Figure 1.



Two obvious ideas would be expressed by “linear” relations in these data. One would be the idea that each additional year of education corresponds to a certain increase in average income, with the number of dollars per year appearing as the slope of the linear relation. The other idea would apply to a linear relation between education and the logarithm of income. In this case each additional year of education would correspond to a multiplication of the average income, a multiple that could also be expressed as a percentage.

While, in principle, it is not logically possible for both of these equations to be correct, the “noise” represented by the vertical scatter in either graph (either income by education or log income by education), is so great that it is not possible to choose between the two possibilities with these data.

I'm also still worrying about the well-behaved or not-well-behaved nature of these variables because it will affect the validity of using any linear technique. So I am going to pull out another property of well-behaved variables, namely that two well-behaved variables should be linearly related (if they are related at all). I'm going to use r to measure the strength of the linear correlation and see whether any combination of transformations has a useful effect on linearity (from which I will infer that I have found the well-behaved transformation).

Correlations	educ	sqrt educ	ln educ
inc	0.38	0.36	0.34
sqrt inc	0.38	0.37	0.35
ln inc	0.34	0.33	0.32

Squared Correlations	educ	sqrt educ	ln educ
inc	0.14	0.13	0.11
sqrt inc	0.15	0.14	0.12
ln inc	0.12	0.11	0.10

That tells me very little: Correlations with sqrt of income are slightly larger than others. Correlations with education are slightly larger than others. But in terms of "variance explained", the range is from 10 percent to 15 percent. And the most interesting fact is that all of them are low: Step away from the methodology: Using these data, at most "15 percent of the variation in income is predictable from education". This is an approximation to one of those facts of life that "everyone" knows to be true. Well, "approximately 15 percent of the variation in income is predictable from education"

Getting ready for an interim report, I will collect various means, and standard deviations, and then write what I can

	Income	Ln Income	Sqrt	Education	Ln	Sqrt
			Income		Education	Education
Mean	$26,473	149.14	9.78	13.69	3.68	2.59

Standard Deviation	$22,295	65.05	1.05	2.75	0.38	0.22

So tentatively, I write what I can — definitely not ready for an external report. Definitely not ready to be the basis of any policy recommendations.

The NORC General Social Survey for 1993 interviewed 1607 individuals selected from a national sample of households	Who, what, where, but also discretely noting my discomfort: I want to talk about individuals but I know that I can't. Yet I am not sure how to translate the findings for this sample, based on households, to a correct statement about individuals.
Respondents showed a mean education of __ years and a mean income of ___ .	Safe, but qualified because I said respondents. (I did not say that the mean education in the US was __ or that the mean income in the US was ___.
While there is a clear correlation between education and income, it is not strong. Estimating the regression line for income as a function of education, the line shows estimates that individuals with no formal education will have __ income while, on the average, each additional year of education corresponds to ___	Using the regression line
In these data, while there is a correlation for the population as a whole, it is not a relation that individuals can count on.	Trying to give some idea of the strength of the correlation

For example, the median income of individuals completing high school was $18,750 in these data , indicating that 50% of these individuals have incomes at or exceeding this amount, 50% have incomes below $18,750 By comparison, individuals with four more years of education, usually college graduates, have a median income of $27,500 but 25 percent of these college graduates had incomes below the median for high school graduates, while 30 percent of the high school graduates had incomes exceeding the median for college graduates. That is, while higher education corresponds to higher income, on the average, there is no guarantee that a specific individual with higher education will have an income exceeding that of another individual with less education.	First try at an illustration. A pair of histograms, one for the income of high school graduates, one for the income of college graduates — both drawn to the same scale — should make this clear.
If we treat income and education together as indicators of social status, the high quartiles of these status variables pair off __ years of education with __ years of income, while the low quartiles of the status distribution pair off __ years of education with __ years of income, approximately __ dollars per year	Implicit use of the orthogonal least squares line. I am showing the correspondence between income and education without using words that imply that one variable depends on the other.

The indications from these data are suspect, requiring some verification before they are used as the basis for other work. First the income and education distributions need to be checked against US Census data, with a clear definition of just who's incomes we are talking about. (Are people who were unemployed and had no other income excluded from the data or were they recorded as zero income.)

The fact that this income distribution did not respond to a logarithmic transformation, by becoming bell shaped and symmetrical, leaves some suspicion of bias such that higher income individuals are less likely to be represented in these data. If so, it will have thrown off all the estimates of correspondence between income and education, probably reducing the income numbers below their true values.

Clear doubts, plus specific reason for the doubts, and anticipation of specific consequences.

The best fit regression equation in dollars and years of education predicts about $3,000 dollars per year of education completed, while the regression equation in log dollars predicts a 14% increase in income per year of education completed. The fact that these two numbers are not consistent points to the large variance of income in all educational groups, making it impossible to be more precise reasoning solely from these data. $3,000 is a little under 14% of the average of $26,500 For a non average education or income it is difficult to be more precise from these data. To give a bench mark to the size of these errors: If a forecaster were to predict an average income, predicting the same income for all persons, ignoring education, the standard deviation of the errors would be $22,000. Using education to improve the prediction of income, the standard deviation of the regression errors would remain at about $20,500.

Nevertheless, the sense of the log equation is a more credible result. It predicts $3,000 as the average income of persons with no education. By contrast, the equation in dollars predicts negative income, $-14,742 for that same person.

Trying to get comfortable with the results.

Similarly, the implications of the error statistics also favor logarithms. The correlation of __ indicates that the variance of the residuals of income is __ percent of __ the variance of income. Where the variance of income is ___, this implies that the standard deviation of the residuals would be ___. Such a number gets us deeper into nonsense: Predicting a standard deviation of ___ among persons with education of 0 and predicted income of ___, a standard deviation of ___ among persons with twelve years of education and predicted income of ___, and the same standard deviation of ___ among persons with twenty years of education and predicted income of ___.

By contrast, using logs, the equation predicts a geometric

In summary, on the average each year of education corresponds to about $3,000 or 14% greater income. But the ranges of income at all educational levels requires a warning that there will be extremely large differences, both positive and negative among individuals with the same education and, therefore, between individuals with different educational backgrounds.	I would not want to leave the reader with the previous paragraph as a final comment.

Re-work the spread sheet: First show the sorting of the data such that identical cases lie in adjacent rows. Then count these cases, displaying the result as a cross-tab. Then modify the spread sheet to compute mean squared error using frequencies. That should get the same job done as r and r-squared which I used in the previous draft. In this draft I can't use r-squared because I haven't prepared the background.

Modify after the work described above:

For the relation in dollars, the coefficient of linear correlation is .38 ; for the relation in log dollars, the coefficient of linear correlation is .34, “explaining” respectively, 14% or 12% of the variation in income.

Where the equation in dollars predicts $3,013 dollars per year of education completed while the equation in log dollars predicts a 14% increase in income per year of education completed. While the log equation produces a more obviously sensible intercept, $3,000 as the average income of persons with no education, it is not altogether obvious that the negative intercept for the dollar equation, $-14,742, is altogether foolish, producing a “deficit” of $15,000. Which should I use? The data are not strong enough to choose for me: I have a prejudice for using logs when I’ve got income data. that would lead me to the log equation. But, I also have a prejudice for sticking with common units if the data do not compel me to do otherwise. That would lead to dollars. But, then again, using dollars only starts out by looking like common sense, and then it presents me with the need to talk about negative income, which is possible but no longer consistent with my prejudice to keep it simple. So I will have to conclude that the data show about $3,000 per year of education and about 14% per year of education. The fact that these two numbers are not consistent points to the large variance of income in all educational groups making it impossible to be more precise reasoning solely from these data. $3,000 is a little under 14% of the average of $26,500 For a non average education or income it is difficult to be more precise from these data. To give a bench mark to the size of these errors: If a forecaster were to predict an average income, predicting the same income for all persons regardless of education, the standard deviation of the errors would be $22,000. Using education, the standard deviation of the regression errors would remain about $20,500.

Return on Investment

Data analysis is fully capable of solving all problems. It is incumbent, therefore, that the skilled analyst choose problems of the highest moral, ethical and scientific value. You can not squander your skills on anything less. Occasionally, however, it is necessary to teach the stuff and demonstrate its power by tackling showy, if less important matters — like getting fabulously rich. I'll start the exercise. Completion is an exercise for the reader.

The data are provided in Fortune's annual report of statistics for the "Fortune 500". I intend to find the variables that predicted the return I can get from investing my money. I will invest my money accordingly and then relax for a few years while my money multiplies. Then I will relax a great deal more.

First, a look at the variables. General Motors' data is shown but for each company the data include

Revenues (Rank in 1995 and 1994)

The *name* of the company,

Revenues (In millions of dollars and as the percent by which revenues changed as compared to 1994.

Profit (in millions of dollars) and as change compared to 1994

Assets (in millions of dollars for 1996 and as an annualized (***?of increase***) during the preceding decade.

Total Stockholders' Equity (in millions of dollars)

Market Value (in millions of dollars)

Profits, shown as
percent of sales
percent of assets
and as Percent of Stockholders' Equity

Earnings per share
in dollars
as percent increase compared to 1994
and as an annualized rate of increase for the decade.

Total Return to Investors as
percentage for 1995 and as
annualized percentage for the decade

The *number of employees*

The *Industry*

And, finally, the address, and the name of the CEO.

First, to preview the wealth I am about to attain: How well could I have done in 1995? To take a look,, I rank order everything according to percentage return for 1995 and I see .... and I see that 10% of the data are outright missing, with no value at all for this variable. So I begin my checklist of things that have to be checked before I can know what to make of any patterns I may find in the 90% of the data that are present.

Now, looking at the numbers, in rank order:

COMPANY	TOTAL RETURN TO INVESTORS			
			1985-95	
	1995		annual rate	
Name	%	FN	%	FN
Continental Airlines	353.3		—	
Northwest Airlines	223.8		—	
USAir Group	211.8		(8.9)	
Sun Microsystems	157.0		—	
Case	114.3		—	
Student Loan Marketing Assn.	109.5		7.7	
First Interstate Bancorp	108.5		15.6	
CompUSA	107.5		—	
UAL	104.3		13.9	
Seagate Technology	97.9		20.7	

Kmart	(42.3)		(0.8)	
Jefferson Smurfit	(44.1)		—	
Merisel	(45.3)		—	
Yellow	(46.9)		(5.5)	
Best Buy	(48.0)		17.6	
Payless Cashways	(54.1)		—	
Flagstar	(55.4)		—	
Penn Traffic	(60.5)		—	
Morrison Knudsen	(64.2)		(12.5)	
Caldor	(85.4)		—	

I see 353.3 percent for Continental Airlines, not bad, I can accept that rate of return. The top ten takes me down to a mere 97.9 percent, still not bad. And I notice a very promising regularity in the names: Continental Airlines, Northwest Airlines, USAir Group, and UAL. Unfortunately, in the cases where there is comparable data for the decade, the rates are not at all this high. Some (in parentheses) are negative. I also note that there is another end to this distribution, Caldor at the extreme (negative 85.4%). Perhaps some caution is in order.

This will have its own problems, but for the beginning of this project I think I will switch to returns annualized over the decade.

Now I see a lot more missing data, 119 of the “500”, and looking at the top and bottom of the list (of those for which there are data)

COMPANY	TOTAL RETURN TO INVESTORS		INDUSTRY
		1985-95	
	1995	annual rate	
Name	%	%	
Home Depot	4.3	44.5	Specialist retailers
Conseco	45.8	42.6	Ins: life & health (stock)
Applied Materials	86.4	40.7	Electronics, electrical equipment
United HealthCare	45.0	38.3	Health care
Micron Technology	80.3	37.2	Electronics, electrical equipment
Nike	88.8	36.8	Wholesalers
Compaq Computer	21.5	36.1	Computers, office equipment
Computer Associates Intl.	76.5	35.1	Computer and data services
Fed. Natl. Mortgage Assn.	75.0	33.1	Diversified financials
Gillette	41.1	31.2	Metal products
Beverly Enterprises	(26.1)	(4.9)	Health care
Yellow	(46.9)	(5.5)	Trucking
Advanced Micro Devices	(33.6)	(5.5)	Electronics, electrical equipment
AST Research	(41.9)	(5.8)	Computers, office equipment
Turner Corp.	1.5	(7.3)	Engineering, construction
PriceCostco 11	18.4	(7.6)	Specialist retailers
USAir Group	211.8	(8.9)	Airlines
Unisys	(36.2)	(10.9)	Computer and data services
Morrison Knudsen 39	(64.2)	(12.5)	Engineering, construction
Navistar International 6	(29.8)	(18.8)	Motor vehicles and parts

That is more believable, 30 to 40% per year, slightly more than two years to double (excluding taxes). Still not bad. There is also some greater reliability of prediction from decade statistics to year statistics, more than the other way around: The top 10 for the decade did well for the year (but not the other way around) — among those companies that survived the decade (and made it to the list). The list shows computing or electronics and health care among both the big winners and big losers.

Now, is the annualized rate of return a well behaved variable? This is a little troublesome. First I'm worried because this is certainly not a random sample of corporations, this is the Fortune "500", the largest (by assets). Moreover, the worst of the lot have probably disappeared, at least from the Fortune 500 and some of the worst and some of the best will have been acquired by others on the list. Whatever that did to the companies, it might well have affected the statistics. In addition, , just dealing with percentages is troublesome. The intervals of a percentage imply that -20%, -10%, 0%, +10%, +20%, +30%, etc. are equal intervals. But percentages stand for simplified ratios. And as ratios, the ratio of .8 to .9 is the same as the ratio of .9 is to 1.01. As ratios, the equal steps are .8, .9, 1.01, 1.14, 1.28, 1.44.

Equal steps as percentages

.8 .9 1.0 1.1 1.2 1.3

Equal steps as ratios

.8 .9 1.01 1.14 .128 1.44

So what are the equal steps, even among percentages? This is a mess, particularly when I know that these particular ratios are annualized ratios out of a ten year span, which means that they have been treated as ratios and then, only for presentation, converted to percentages. So, if what I want is to represent data computed as ratios, and to represent them in a form that represents equal ratios as equal intervals, then I will use the logs of the ratios.

I like that argument and, with confidence I was about to show off by "discovering" that these numbers were well-behaved once converted to logs. Trouble is, they are not. In logs these returns have a skewed distribution with a tail to the left. (The sequence of mid values, from the median value to the mid thirty-second value, decreases.)

Count	Value	Value	Mid Value	
n=381				
191	0.130	0.130	0.130	Median
96	0.159	0.094	0.127	Mid Quartile
49.5	0.188	0.057	0.123	Mid Eighth
25	0.225	0.014	0.119	Mid Sixteenth

13	0.250	-0.028	0.111	Mid Thirty-Second
7	0.308	-0.060	0.124	Mid Sixty-Fourth

Because it violates my naive expectations, this would be a bit disconcerting, except that all the other complications affecting this distribution suggest that I give it a little leeway — if not the logs, then something close.

So, back to the numbers. Square roots. Still skewed to the left.

Using the original numbers. Still skewed to the left. Now I've given it more than a little "leeway". Still not well behaved.

Count	Value	Value	Mid Value	
n=381				
191	1.139	1.139	1.139	Median
96	1.172	1.099	1.136	Mid Quartile
49.5	1.207	1.059	1.133	Mid Eighth
25	1.252	1.014	1.133	Mid Sixteenth
13	1.284	0.972	1.128	Mid Thirty-Second
7	1.361	0.942	1.152	Mid Sixty-Fourth

So blindly, at least for the moment, I have to ask just what it would take to make this distribution symmetrical. And this is going in exactly the opposite direction from what I expected,

not

1, .5, 0 (no transform, square root, log,,, decreasing)

but

1, 2, 3 ... (no transform, square, cube, increasing)

Proceeding, it takes something like a cube to get the median, mid-quartile, and mid-eighth to line up.

Count	Value	Value	Mid Value	
n=381				
191	1.478	1.478	1.478	Median
96	1.610	1.327	1.469	Mid Quartile
49.5	1.756	1.188	1.472	Mid Eighth
25	1.963	1.043	1.503	Mid Sixteenth
13	2.117	0.918	1.518	Mid Thirty-Second
7	2.521	0.836	1.678	Mid Sixty-Fourth

That is interesting. It violates my naive expectation, quite sharply, so I have to ask why — and maybe learn something.

I don’t take the detail seriously, not the third power, as compared to the 2.5th power or the 2nd power. But I do take the direction leading to the 2nd or 3rd power seriously. What does it mean?

Consider the opposite. Suppose that logs had worked. In dollars, that would mean that each additional dollar is easier. One dollar profit on one dollar is hard, one dollar profit on one thousand dollars is easy.

This is running in the opposite direction. It means that once you have made a 1% return, the next 1% is "harder", and the next harder, or less likely. Going down it means that going down 1% is easier if you've already dropped a couple of percent. In stock market terms it means that going up is more likely to be slow, going down more likely to be sudden. Don't trust this — because I'm reaching in my mind for models or "folklore" to match the clue provided by the data — but that is what I read as a strong possibility, based on the shape of the distribution. In value terms, when a value is dropping, the chances increase that it will drop precipitously (get out fast). When a value is rising, the chances that it will continue to rise diminish.

Consider the implication. There is no proof here, nothing more than these numbers. But consider the implication to see the issue raised by this peculiar distribution. There is a minor industry organized to provide information to investors, very often graphical information. People look at these things. And although "technical analysis" of the stock market gives names and numbers to patterns, first people look at the numbers. Then they try to formalize patterns that may, or may not, be present.

To "look" at these numbers, charts are frequently provided on two scales:

One is the price of the stock. The alternative is the logarithm of the price of the stock — done for all the common sense reasons that we have already discussed about logarithms — constant rates of change (straight lines) on such graphs correspond to constant rates of increase at continuously compounded rates. This evidence, suggests that the place to look for simplicity lies in quite another direction, the squares or cubes, not the logs. Again — that is a lot of guessing, and at least a bit of sloppy thinking. (Note I slipped from ratios, which are a measure of change, to dollars which are not.) I do not submit this to you, or to myself as a fact, as a conclusion, or even as an inference. I am already criticizing it, reversing it, and generally turning it around as I think about it. I'm letting you see the usually hidden process of reasoning and guessing. This is what follows from looking at the shape of this distribution, finding a surprise, and thinking about it.

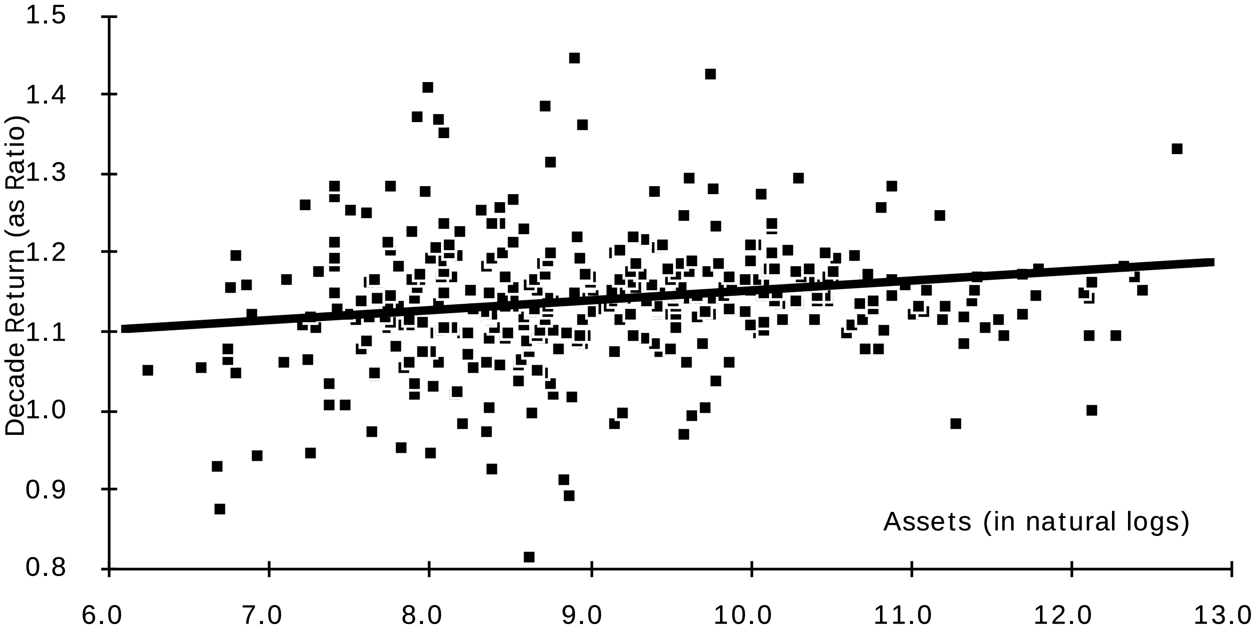
Can I predict these returns? For that I need some correlation with other variables, preferably a strong correlation. I'll begin with assets. Is it true that as corporations get larger, their rates of growth must diminish? Is it true that to find large returns I have to look to middle size, or smaller corporations.

The assets variable is surely truncated — the selection criterion for the Fortune list selected the largest corporations. So the first criterion for a well behaved variable can not be examined. Homeoscedasticity is a possibility. But I am going to do a quick start with logs and go directly to the relation, if any, between assets and return.

Using my spread sheet program, the number I get for the correlation is .14, and the number I get using the more exotic third power is .__. That tells me not to worry a great deal about that transformation: Where most of the data lie between .95 and 1.3, curvature within this range is not going to be difficult to distinguish from a straight line.

Now that correlation tells me that approximately 2% of the variance in return is predictable from log assets. Can I be seriously interested in something that “explains” 2% of the variance in the variable I am trying to predict?

It all depends on context. And in this case I am very interested. Look at the graph of the data together with the regression line. It has an intercept of 1.0524 and a slope of .0092 (with a correlation of .145).



Just looking at the graph, there are lots of corporations to choose from with log assets approximately 7.5 (assets approximately 1.8 *billion* dollars). And there are lots of corporations to choose from with log assets approximately 10.5 (assets approximately 36.3 billion dollars). With a slope of .0092, the expected difference between their average returns (as ratios) is

.0092 times 3 = .0276

That means that on the average the higher group realized 2.76% greater return. The expected return for the higher group was 1.1491, as a ratio; the expected return for the lower group was 1.1215, as a ratio. Or, as percentages 15% versus 12% return on the investors' dollar. The variation around this return is enormous:

Reading it directly off the graph, it is common to find values between the predicted ratio -.15 and the predicted ratio +.15.

Computing it :

The mean return, as a ratio, is 1.136, as a percentage that is 13.6% annual return to the investor.

The standard deviation of the ratio is .0805; the variance of the ratio is .0065.

In standardized form the variance of the residuals is root(10r-squared) is .9895, 98.95% unexplained.

Knowing that the variance of the return is .0065, that leaves the ratio with an error variance of .0064

and an error standard deviation of .0801, which is the square root of this number.

So, two standard deviations of error above and below corresponds to the predicted ratio ±.1602 — very close to my eyeball estimate.

Can I find anything useful in such small correlations and large variance? You bet, literally, you *bet*. I can bet my money on 20 or 30 securities in the group with log assets of about 7.5 and get a return fairly close to the average for that group — and the average for that group is about 3% on the dollar greater than the lower asset group. Taking not those with assets of 7.5 in logs but taking the largest, and comparing their expected value to the mean return for all these investments, I should clear maybe 3% more than the average. And in the investment, beating the averages by 3%, is doing very well, gaining perhaps 16 or 17% per year rather than the average 13.5% per year.

So yes, this r is tiny. But for investment purposes my primary interest is in the slope. And the size of this slope is quite usable. Then, knowing that the r is small, I will have to protect myself by diversifying, in order to protect myself from the variance and realize the gains predicted by the slope.

So yes, I can work with these tiny correlations. The problem here is not the correlations. The problem here is that I haven't *really* predicted anything. The words, "prediction", "estimation", "error", and so forth are statistical conventions. But check their meaning operationally, with these data: These numbers tell me how well I could have done, over the decade ending last year if, at the beginning of the decade, I had had the 1995 assets and the regression data that became available at the end of the decade. How well this variable will, or would, predict into the next decade or the next year is an interesting question, not answered by these data — though the exploration of these data provides an interesting lead.

Note on multiple regression

I have barely begun this exploration. Are their variables showing a higher correlation with return? (Yes)

It is also possible to try to predict one variable, Return, by writing a linear equation using two or more predictors, not just one.

$y = a_0 + a_1x_1 + a_2x_2$.

Most statistical programs will accommodate "multivariate regression". The interpretation of multivariate linear equations has some surprises in it, but the basics begin like the basics of two variable regression: Start with well behaved variables. Generate residuals comparing y as predicted from the y that is observed. (you can graph these residuals as the ordinates in two or more different graphs of residuals, one for each of the predicting variables.) Then, if you want a number saying “how good”, compared the variance and standard deviation of the residuals to the variance and standard deviation of the original variable “y”.

Trend

I have data on fat. Not just any fat, but my fat. I'm told that less is better, within certain ranges. It is not at all clear to me how to change body fat or whether to change body fat. But one thing I do know about is gadgets, I like gadgets and data. So one thing I am sure I can do is to convert body fat into is a high tech distraction replete with gadgets and data.

The gadget is a Futrex 1000. I press it into the skin (and fat) above my bicep. It returns a three digit number estimating body fat. According to directions,

> …human bodyfat absorbs light at specific wavelengths in the near-infrared portion of the spectrum. The Futrex 1000 emits these near-infrared wavelengths. Also, the Futrex-1000 contains an optical sensor to measure how much of this energy is absorbed by your bodyfat.
>
> The near-infrared absorption technology used in the Futrex-1000 has been shown to be well within the accuracy of ±4.5%. Additionally, due to the smaller number of test variables, the repeatability of near-infrared is superior to all of the other bodyfat determination methods. …
>
> …Everyone's body responds to diet and exercise differently. For example, the "average person" who loses one pound of weight due to dieting, actually loses approximately 3/4 of a pound of fat and 1/4 pound of lean. However, if that same "average person" loses one pound of body weight due to exercise, he actually loses approximately 1 1/4 pound of fat and gains 1/4 pound of lean (muscle). Thus the benefit of exercise is obvious.
>
> You should be aware that when you start an exercise program there will not be an immediate reduction in your percent bodyfat. This is because the body will first lose water. As you continue your program, you will begin to lose bodyfat. It usually takes several weeks before noticeable loss in percent bodyfat occurs.

That tells me, among other things, that the measurement includes error. The magnitude is, it tells me "within the accuracy of ±4.5%.

It also tells me that it is measuring something other than body fat. It hints at that in the little discourse on losing water. And, whether the directions told me so or not, with most "measurements" are not direct measurement of the thing they claim to measure: It is measuring absorption of light, not fat. That means that the numbers on its face are connected to the true (but unknown) facts for my body by a combination of theory and indirect physical links that connect infrared to fat. Every step in the process is an opening to sources of error. Every step in the process opens the door to other variables which can affect the measurement. These are the variables referred to in the famous caveat "all things being equal".

So, Wednesday, July 3 1:30 in the afternoon. I am going to reduce uncertainty by using repeated measurements. The ten estimates are

24.8%	25.6%	26.4%	24.2%	26.2%	26.1%	24.2%	23.4%	23.2%	25.8%

I'm not sure what the directions referred to as ±4.5%. Was that +4.5% of the observed value? Or was it 4.5% added or subtracted from the observed value? I don't know. But the range in these ten observations is 3.2%, from a low of 23.2% to a high of 26.4%. The mean is 25.0%. The standard deviation is 1.2%. Plus or minus two standard deviations would be plus or minus 2.4%, disconcertingly different from the 4.5% number specified in the directions.

That difference alerts me to another inevitable fact: I have to distinguish between variation around the mean (measured by the standard deviation), and variation of the mean (and the whole distribution) around the true value. One is the variation around the measurement, all things being equal. The other is the variation of the measurements induced by the fact that all things will not be equal from day to day.

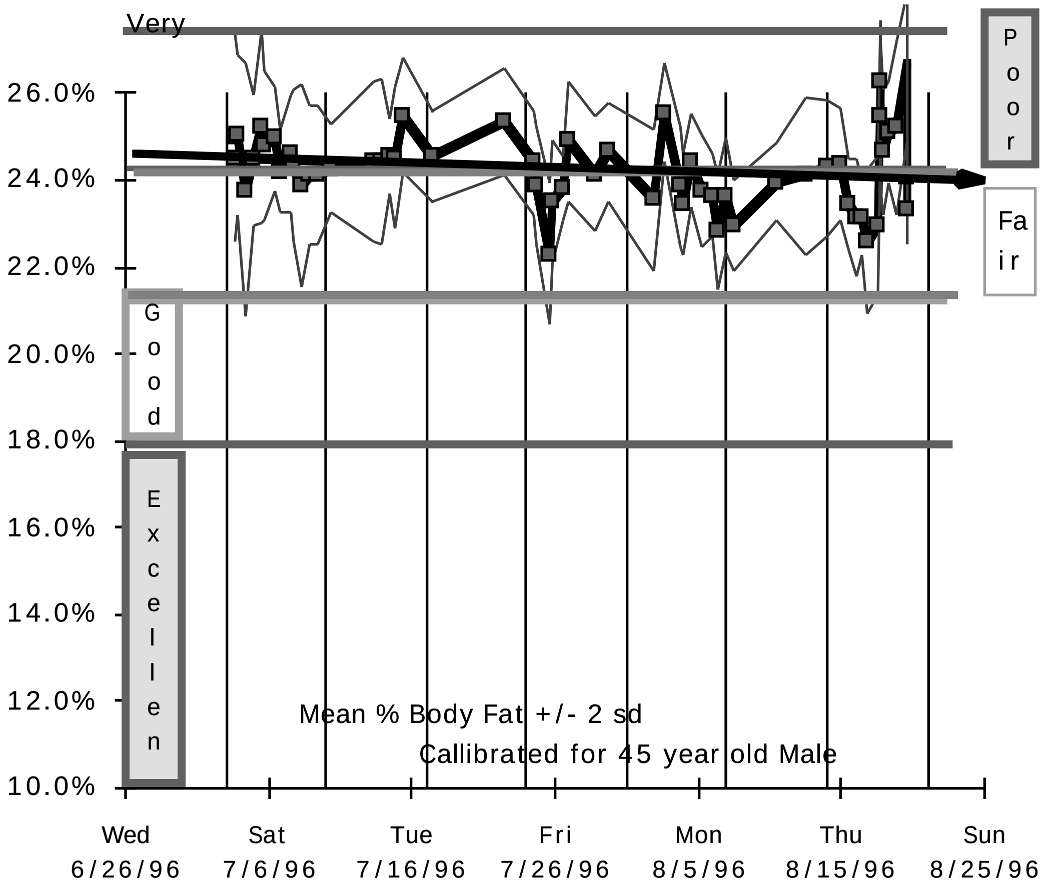
If the measurements themselves have 4.5% error built in (whatever that means) can I detect a trend within all this variation and all this error? With 5 or 6 weeks of data, 1 or 2% change is the most that would be credible since I have neither eliminated food nor dedicated myself exclusively to exercise.

Here is the graph and here are the data. They are not equally spaced in time. I did not always take ten measurements. And toward the end I took many measurements because I couldn't believe the numbers. The last measurements are close together because

I was trying to figure out what could possibly be going on that I would (appear to have) gained, in one instance, 3% during one hour. So I took multiple measurements, hoping to outweigh the bizarre values with others that would return to what had appeared normal. The result was the opposite, replicating the bizarre values.

So, is there a trend? How much? (The annotations, Very Poor, Poor, Fair, and so forth, are based on the instruction manual.

I presume that you can collect and analyze analogous data for a variety of physiological measures and indicators of performance.



Date	Mean Percent Body Fat	Standard Deviation										
Wed 7/3/96 13:42	25.0%	1.2%	24.8%	25.6%	26.4%	24.2%	26.2%	26.1%	24.2%	23.4%	23.2%	25.8%
Wed 7/3/96 17:42	25.0%	0.9%	24.3%	25.5%	25.0%	24.1%	23.8%	25.2%	25.7%	26.7%	25.8%	24.3%
Thu 7/4/96 11:02	23.8%	1.5%	23.5%	23.0%	24.7%	22.6%	27.4%	22.5%	22.8%	23.6%	23.4%	24.2%
Thu 7/4/96 21:00	24.5%	0.8%	23.7%	23.5%	24.2%	24.8%	25.2%	23.9%	24.6%	23.8%	25.6%	25.4%
Fri 7/5/96 11:00	25.2%	1.1%	24.4%	23.5%	24.3%	25.6%	24.5%	24.8%	26.0%	26.9%	26.7%	25.4%
Fri 7/5/96 17:00	24.8%	0.9%	25.5%	25.9%	24.8%	25.0%	25.8%	24.5%	25.0%	23.7%	23.2%	24.5%
Sat 7/6/96 9:00	25.0%	0.6%	25.0%	25.3%	24.6%	25.9%	25.0%	25.5%	23.9%	25.1%	24.2%	25.1%
Sat 7/6/96 18:00	24.2%	0.5%	24.1%	23.4%	23.6%	24.4%	24.7%	24.9%	24.5%	24.3%	24.3%	23.8%
Sun 7/7/96 11:00	24.6%	0.7%	24.6%	25.7%	24.5%	23.8%	24.3%	23.9%		24.7%		25.5%
Sun 7/7/96 16:00	24.3%	0.9%	25.5%	24.6%	25.5%	23.4%	22.8%	24.5%	24.5%	23.6%	24.1%	24.7%
Mon 7/8/96 9:00	23.9%	1.2%	24.5%	22.9%	24.2%	24.0%	23.8%	21.9%	22.4%	25.4%	24.0%	25.4%
Mon 7/8/96 21:00	24.1%	0.8%	24.0%	25.2%	25.7%	24.0%	23.2%	23.2%	23.6%	23.8%	24.5%	24.0%
Tue 7/9/96 9:00	24.1%	0.8%	24.0%	25.2%	25.7%	24.0%	23.2%	23.2%	23.6%	23.8%	24.5%	24.0%
Wed 7/10/96 9:00	24.3%	0.5%	23.9%	24.5%	24.3%	24.1%	24.8%	23.1%	24.6%	24.7%	24.5%	24.0%
Sat 7/13/96 9:00	24.4%	0.9%	25.8%	24.2%	24.6%	22.9%	24.9%	24.2%	25.4%	23.4%	25.1%	23.7%
Sat 7/13/96 21:00	24.4%	1.0%	22.3%	24.1%	24.9%	24.3%	24.3%	24.5%	25.5%	25.6%	25.0%	23.8%
Sun 7/14/96 9:00	24.6%	0.4%	24.4%	25.3%	24.1%	24.7%	24.5%	25.0%	24.6%	23.8%	24.2%	24.8%
Sun 7/14/96 21:00	24.5%	0.8%	26.2%	23.7%	25.2%	24.5%	24.5%	23.3%	23.9%	24.9%	24.3%	24.6%
Mon 7/15/96 9:00	25.5%	0.6%	24.2%	26.2%	25.0%	24.4%	25.7%	25.5%	25.7%	25.9%	25.5%	25.0%
Wed 7/17/96 9:00	24.6%	0.5%	24.2%	24.5%	23.6%	24.7%	24.6%	25.4%	24.9%	24.7%	23.8%	24.5%
Wed 7/17/96 9:00	24.6%	0.5%	24.2%	24.5%	23.6%	24.7%	24.6%	25.4%	24.9%	24.7%	23.8%	24.5%
Mon 7/22/96 9:00	25.4%	0.6%	25.1%	25.2%	25.5%	25.8%	24.1%	25.6%				
Wed 7/24/96 10:00	24.4%	0.6%	24.1%	24.5%	24.3%	25.6%	23.9%	24.5%				
Wed 7/24/96 16:00	23.9%	0.7%	22.3%	24.1%	23.8%	24.0%	23.6%	24.0%				
Thu 7/25/96 16:00	22.3%	0.8%	20.5%	21.9%	22.1%	23.3%	22.6%	23.0%	21.8%	22.5%	21.6%	22.9%
Thu 7/25/96 21:00	23.5%	0.7%	24.5%	23.8%	23.2%	23.5%	24.0%	23.3%	21.8%	23.3%	23.5%	23.8%
Fri 7/26/96 12:00	23.8%	0.4%	24.2%	23.8%	22.9%	23.7%	24.2%	23.8%	23.7%	24.0%	23.9%	23.7%
Fri 7/26/96 20:00	24.9%	0.7%	25.0%	25.1%	25.4%	24.1%	23.2%	25.0%	25.5%	24.8%	24.4%	24.6%
Sun 7/28/96 18:00	24.2%	0.7%	23.3%	23.8%	23.3%	24.5%	24.3%	25.1%	24.0%	25.2%	23.8%	24.6%
Mon 7/29/96 18:00	24.7%	0.6%	23.8%	24.7%	24.5%	24.8%	25.4%	24.7%	24.6%	24.4%	24.2%	25.8%
Thu 8/1/96 23:00	23.6%	0.8%	25.5%	23.4%	24.6%	23.6%	23.5%	22.7%	24.0%	24.5%	23.3%	23.4%
Fri 8/2/96 15:00	25.6%	0.6%	25.6%	24.8%	25.9%	24.4%	26.1%	25.5%	25.8%	25.2%	26.0%	24.9%
Sat 8/3/96 18:00	23.9%	0.7%	24.5%	23.3%	23.1%	24.3%	25.4%	23.2%	23.8%	23.9%	23.6%	24.1%

Sat 8/3/96 21:00	23.5%	0.6%	24.0%	23.9%	22.3%	23.2%	24.3%	23.5%	22.9%	23.4%	23.6%	23.0%
Sun 8/4/96 13:00	24.5%	0.5%	24.2%	25.1%	24.2%	25.5%	25.0%	23.7%	24.2%	24.2%	24.7%	24.7%
Mon 8/5/96 8:00	23.8%	0.6%	24.3%	23.4%	22.9%	24.1%	23.6%	22.8%	24.3%	23.1%	24.6%	23.9%
Mon 8/5/96 22:00	23.7%	0.5%	23.7%	22.9%	23.7%	24.6%	23.3%	23.6%	23.5%	23.1%	23.9%	23.9%
Tue 8/6/96 10:00	22.8%	0.7%	23.0%	23.9%	22.1%	22.3%	22.1%	23.3%	22.0%	23.6%	22.8%	22.8%
Tue 8/6/96 22:00	23.7%	0.7%	23.2%	23.6%	24.1%	22.2%	22.5%	23.9%	23.8%	24.0%	22.9%	23.7%
Wed 8/7/96 10:00	23.0%	0.5%	22.8%	24.2%	22.8%	22.1%	22.9%	23.2%	23.0%	22.6%	23.0%	23.1%
Sat 8/10/96 13:00	24.0%	0.5%	24.6%	24.1%	23.4%	24.5%	24.2%	24.0%	23.4%	23.6%	23.4%	23.9%
Mon 8/12/96 13:00	24.1%	0.9%	24.2%	25.0%	23.9%	23.6%	22.0%	24.4%	24.4%	25.4%	23.9%	24.0%
Tue 8/13/96 22:00	24.3%	0.8%	24.0%	25.4%	23.5%	25.3%	25.3%	24.0%	24.3%	23.4%	25.3%	24.3%
Wed 8/14/96 22:00	24.4%	0.6%	23.8%	24.4%	24.3%	25.7%	25.4%	24.5%	25.1%	24.0%	24.1%	24.0%
Thu 8/15/96 10:00	23.5%	0.5%	23.6%	22.9%	23.2%	24.0%	23.6%	23.7%	23.3%	23.2%	24.5%	22.7%
Thu 8/15/96 23:50	23.2%	0.7%	23.1%	23.1%	23.2%	22.9%	23.5%	24.7%	22.4%	23.8%	22.4%	23.6%
Fri 8/16/96 11:50	23.2%	0.4%	23.2%	23.6%	23.5%	23.2%	22.8%	22.8%	23.4%	23.1%	23.1%	22.1%
Fri 8/16/96 19:00	22.6%	0.8%	23.2%	23.6%	24.3%	21.8%	23.5%	22.6%	22.1%	22.6%	22.2%	22.0%
Sat 8/17/96 11:00	23.0%	0.8%	22.8%	23.9%	21.6%	24.2%	23.4%	24.0%	23.1%	22.4%	22.8%	22.6%
Sat 8/17/96 18:00	26.3%	0.7%	25.8%	26.5%	26.5%	27.5%	26.1%	27.2%	25.4%	25.8%	25.7%	26.7%
Sat 8/17/96 19:30	25.5%	0.9%	24.1%	25.2%	26.2%	26.4%	25.6%	25.5%	26.4%	24.4%	24.3%	25.4%
Sat 8/17/96 20:00	24.7%	0.7%	23.6%	24.8%	24.0%	24.3%	26.0%	24.5%	23.8%	25.2%	24.8%	24.9%
Sun 8/18/96 9:00	25.1%	0.6%	25.6%	25.1%	26.1%	26.1%	26.3%	25.1%	25.1%	25.1%	24.7%	24.7%
Sun 8/18/96 21:00	25.2%	1.0%	23.3%	25.2%	22.9%	24.7%	25.5%	25.2%	25.5%	25.3%	25.4%	23.6%
Mon 8/19/96 12:40	26.7%	0.9%	26.7%	27.7%	25.9%							
Mon 8/19/96 12:40	23.3%	0.4%	24.0%	23.5%	23.1%	23.1%	23.2%	24.1%	23.5%	23.4%	23.1%	22.9%

. and if, as I can almost guarantee, I have not lost fat on that order of magnitude during the first month or two, can I even detect a trend in the measurements of that time. (I presume a drop from 25% bodyfat on July 3 to 20% body fat in the middle of August would be quite noticeable: If it were pure loss, without increase in muscle, that would be 7 or 8 pounds in 5 or 6 weeks. I would know if something would be quite dramatic if it occurred

Mean-Based Budgeting

My grade school textbooks, in Chicago, told me that the model of democracy could be found in the New England town meeting. That was where neighbors came together, argued their visions of the good the true and the beautiful in fair debate, and wrote the laws they would live buy. Displacing heaven to New England saved my grade school teachers from potentially interesting but much less theoretical discussions of the experience we lived with in Chicago.

My adult experience in New Hampshire brings me fact to face with the New England town meeting. And I can say this much for it: If the stuff that teachers feed nine year olds had turned out to be true, then civic life in New England would have been rather dull, which it is not.

The key item on the agenda each year is the budget. There it is: A town of 3,979 people; last year's appropriations, $3,301,133; proposed budget $3,077,903 (not including the school budget). "What is your pleasure on the budget?"

So what do *you* think? Is $3,077,903 a good budget? That's a little hard to answer, isn't it. Well the budget comes to the voters in a booklet disclosing some of the detail. And long before it is brought to the voters, "What is your pleasure on the budget?", it goes through a budget committee, and through Selectmen, and through public hearings.

Would detail help you evaluate the budget? O.K.:

Welfare: Direct Assistance

Actual appropriations prior year:
$44,000

Actual expenditures prior year:
$22,738

Selectmen's Recommended Budget:
$29,000

Budget Committee Recommendation:
$31,500

That doesn't really help much. And I don't recall that the wispy picture of democracy drawn in my Chicago text book offered much guidance for creating budgets, spending real money on welfare, roads, salaries, and so forth.

There is no formula that will translate the grand democratic vision into a budget—ultimately it is a matter of vision and interest, what is needed, who says so, what is right, what is legitimate, who is responsible, what can we afford. More than numbers, a budget debate includes demands from department heads, from affected citizens, from community groups, businesses, and other governments — all of which represent themselves with varying degrees of effectiveness and persuasion.

But the discussion can be given context (followed by argument about which context is relevant). Most easily: What did we spend last year? The context does not establish an ethical calculus that will translate beliefs into numbers. But it helps: Just as the shape of a statistical distribution indicates what is average (without claiming that it is right), the size of last year's budget indicates where the discussion will begin. Just as the extremes of a statistical distribution indicate what is atypical, unusual numbers in a budget will focus the debate.

The advantage of last year's numbers as a context for this year's debate is that the numbers are available. The trouble with last year's numbers as a context is that they are insensitive to deviations that grow by degree: Small deviations accumulate into large excursions, imperceptibly, year by year: One year a department merits an increase 2 or 3 percent in excess of other departments. Next year the department needs a new piece of equipment. The next year the equipment requires maintenance. After that there should be a capital reserve. Next year there is an unrepeatable emergency. After that there is a building on the market at a price which, in the long run, saves money. Each change is reasonable and not excessively different from the year before. And then, year by year, insignificant deviations accumulate into significant distortions — which are undetectable in the comparison between one year and the last. It is like adding a few grains of sugar to a cup of coffee, a few grains at a time. The change is undetectable. But eventually the coffee will surely become too sweet.

A better context can protect against these local excursions into madness by asking not, "What did we do last year?", but asking "What does everyone else do?", where "everyone else" means the voters in other towns in the state. Other towns in the state operate with the same state laws, (governing, for example, whether or not the "town" budget includes the school budget), with the same weather attacking the roads, and with similar economies. In this context the mean provides a base line. And the shape of the distribution and the extremes provide a reality check free of local personalities, free of local credit or blame: "Three quarters of the towns in this state are spending half as much per person on welfare as we are. Why is that?" It gives focus to the discussion.

Until recently mean based budgeting was a good plan that was impossible to implement: You knew the next town. You could look at their books. But budgets were on paper. Budgets followed different accounting categories. It was impractical to develop the empirical base that would support these local decisions. It was impractical to begin with the simple statement

> "On the average, towns spend x-dollars per capita on welfare [using the median]. And fifty percent of the towns voted appropriations between a-dollars per capita and b-dollars per capita [using the two quartiles]."

followed up by the simple question

> "Our town spends 20% more (or 20% less) than the 'normal' range? Why?"

Now, it is practical. For better or for worse, towns have accommodated themselves to computers, states have archived these computerized accounts and imposed common accounting categories. Archives are open to public access.

The statistics are straight forward: If the object is to place the total budget in context, then "regress" total budget as a function of total population. Fit a line to the data, compute the residuals, and that's it: X-burg is __ above or __ below what you would expect for a town of its size. If the objective is to place the welfare budget in context, or the road budget, or the budget for the town office in context: regress the objective on population size, and compare X-burg to the statistical norms for a town of its size.

The statistics are straightforward. But first, there is a question. Is this comparison valid? Is there any reason to believe that budgets either are (or should be) proportional to population? Is there evidence that this statistical criterion matches the problem to which it is being applied?

Folklore and common sense are richly contradictory on this question:

For example, everyone knows that large cities have problems of crime notencountered by small towns. Therefore the per capita cost of public safety will be (or should be) higher in large cities. Everyone knows that you can not compare the budgets of small towns to the budgets or larger cities, even on a per capita basis: Cities cost more.

Common sense, with its rich supply of contradictory advice also tells us that it is easier (more efficient) to large groups than small ones. A town like mine can, on the average, expect about .05 fires at any hour of the day or night, but it still has to have one full fire crew at the ready when the call comes in. A large city can match its capacity much more closely to the demand, lowering the cost per capita. On a per capita basis small towns cost more.

So common sense, as usual, is eloquently useless, able to support any proposition, or deny it, or both confirm it and deny it at the same time. Perhaps the facts can do better: Is it valid to compare costs from town to town, on a per capita basis. Empirically, is there a relation between cost and population? What is that relation? Does it provide a usable base for comparing budgets, town to town?

The Relation Between Population and Total Budget

Beginning with the variable that sums up the rest, beginning with the “bottom line”, here are the population and budget numbers for all towns of the state of New Hampshire in 1994.

TOWNNAME	Population	Log Population (Base 10)	1994 Final Appropriation (Excluding School Budget)	Log 1994 Appropriation (Base10)
HARTS LOCATION	36	1.56	16,000	4.20
ELLSWORTH	74	1.87	30,808	4.49
WINDSOR	107	2.03	58,917	4.77
WATERVILLE VALLEY	151	2.18	1,740,814	6.24
EASTON	223	2.35	90,963	4.96
CLARKSVILLE	232	2.37	145,950	5.16
ORANGE	237	2.37	143,221	5.16
ROXBURY	248	2.39	76,421	4.88
CHATHAM	268	2.43	84,028	4.92
ERROL	292	2.47	211,349	5.33
SHARON	299	2.48	138,500	5.14
GROTON	318	2.50	241,847	5.38
DUMMER	327	2.51	184,018	5.26
BENTON	330	2.52	56,121	4.75
LANDAFF	350	2.54	181,648	5.26
EATON	362	2.56	299,695	5.48
RANDOLPH	371	2.57	221,855	5.35
HEBRON	386	2.59	267,954	5.43
LYMAN	388	2.59	345,987	5.54
DORCHESTER	392	2.59	178,009	5.25
SHELBURNE	437	2.64	302,111	5.48
SUGAR HILL	464	2.67	520,871	5.72
BROOKFIELD	518	2.71	276,403	5.44
STARK	518	2.71	274,850	5.44
CARROLL	528	2.72	569,912	5.76
NELSON	535	2.73	302,008	5.48
ALBANY	536	2.73	447,992	5.65
LANGDON	580	2.76	286,900	5.46
STODDARD	622	2.79	395,507	5.60
PIERMONT	624	2.80	287,212	5.46
CROYDON	627	2.80	292,970	5.47
WASHINGTON	628	2.80	865,975	5.94
WENTWORTH	630	2.80	529,955	5.72
MARLOW	650	2.81	345,725	5.54
COLUMBIA	661	2.82	202,933	5.31
SURRY	667	2.82	233,351	5.37
JACKSON	678	2.83	976,718	5.99
SULLIVAN	706	2.85	256,805	5.41
SOUTH HAMPTON	740	2.87	324,468	5.51
GOSHEN	742	2.87	334,423	5.52
GILSUM	745	2.87	266,667	5.43
MONROE	746	2.87	456,612	5.66
ACWORTH	776	2.89	387,067	5.59
BATH	784	2.89	450,261	5.65
SPRINGFIELD	788	2.90	777,247	5.89
BRIDGEWATER	796	2.90	488,662	5.69
FRANCONIA	811	2.91	742,157	5.87
HILL	814	2.91	421,541	5.62
WARREN	820	2.91	329,435	5.52
DALTON	827	2.92	594,631	5.77
NEW CASTLE	840	2.92	1,067,544	6.03
RICHMOND	877	2.94	327,495	5.52
DANBURY	881	2.94	413,890	5.62
NEWFIELDS	888	2.95	741,162	5.87
PITTSBURG	901	2.95	678,073	5.83
GRAFTON	923	2.97	549,815	5.74
STRATFORD	927	2.97	1,443,059	6.16
FREEDOM	935	2.97	1,067,095	6.03
WILMOT	935	2.97	513,349	5.71
EFFINGHAM	941	2.97	664,421	5.82
LEMPSTER	947	2.98	504,417	5.70
JEFFERSON	965	2.98	375,173	5.57
HARRISVILLE	981	2.99	546,486	5.74
NEWINGTON	990	3.00	2,865,784	6.46
CENTER HARBOR	996	3.00	828,850	5.92
ORFORD	1,008	3.00	731,668	5.86
STEWARTSTOWN	1,048	3.02	429,166	5.63
SALISBURY	1,061	3.03	515,522	5.71
SANDWICH	1,066	3.03	1,451,472	6.16
WOODSTOCK	1,167	3.07	1,564,914	6.19
MIDDLETON	1,183	3.07	521,861	5.72
ALEXANDRIA	1,190	3.08	717,376	5.86
TEMPLE	1,194	3.08	606,798	5.78
MASON	1,212	3.08	675,578	5.83
FRANCESTOWN	1,217	3.09	960,260	5.98
LINCOLN	1,229	3.09	2,960,520	6.47
BENNINGTON	1,236	3.09	864,319	5.94
GRANTHAM	1,247	3.10	872,350	5.94
LYNDEBOROUGH	1,294	3.11	694,035	5.84
MILAN	1,295	3.11	435,206	5.64
UNITY	1,341	3.13	575,507	5.76
NEWBURY	1,347	3.13	1,216,672	6.09
EAST KINGSTON	1,352	3.13	641,087	5.81

MADBURY	1,404	3.15	670,873	5.83
BRADFORD	1,405	3.15	1,095,186	6.04
WEBSTER	1,405	3.15	696,631	5.84
RUMNEY	1,446	3.16	533,389	5.73
SUTTON	1,457	3.16	1,030,573	6.01
DUBLIN	1,474	3.17	938,545	5.97
LYME	1,496	3.17	1,209,899	6.08
HAMPTON FALLS	1,503	3.18	912,400	5.96
THORNTON	1,505	3.18	1,371,051	6.14
GREENFIELD	1,519	3.18	672,559	5.83
WESTMORELAND	1,596	3.20	410,500	5.61
HANCOCK	1,604	3.21	1,076,495	6.03
NEW HAMPTON	1,606	3.21	1,264,561	6.10
KENSINGTON	1,631	3.21	581,154	5.76
CORNISH	1,659	3.22	895,466	5.95
LISBON	1,664	3.22	1,221,587	6.09
CANTERBURY	1,687	3.23	615,537	5.79
HOLDERNESS	1,694	3.23	1,498,571	6.18
MADISON	1,704	3.23	1,198,574	6.08
DEERING	1,707	3.23	862,174	5.94
ALSTEAD	1,721	3.24	765,745	5.88
DUNBARTON	1,759	3.25	664,876	5.82
MONT VERNON	1,812	3.26	863,697	5.94
TUFTONBORO	1,842	3.27	983,087	5.99
ANDOVER	1,883	3.27	562,335	5.75
WHITEFIELD	1,909	3.28	2,011,131	6.30
ASHLAND	1,915	3.28	5,991,549	6.78
MARLBOROUGH	1,927	3.28	1,484,826	6.17
CHICHESTER	1,942	3.29	696,019	5.84
NEW DURHAM	1,974	3.30	1,709,885	6.23
FITZWILLIAM	2,011	3.30	1,105,147	6.04
BETHLEHEM	2,033	3.31	1,695,314	6.23
PLAINFIELD	2,056	3.31	1,252,835	6.10
TROY	2,097	3.32	1,174,027	6.07
SANBORNTON	2,136	3.33	1,485,044	6.17
TAMWORTH	2,165	3.34	1,159,218	6.06
GREENVILLE	2,231	3.35	1,303,897	6.12
WARNER	2,250	3.35	1,513,819	6.18
BARTLETT	2,290	3.36	1,143,921	6.06
ANTRIM	2,360	3.37	1,688,818	6.23
CAMPTON	2,377	3.38	1,040,835	6.02
BROOKLINE	2,410	3.38	1,113,194	6.05
COLEBROOK	2,444	3.39	1,357,725	6.13
NORTHUMBERLAND	2,492	3.40	1,371,920	6.14
DANVILLE	2,534	3.40	858,789	5.93
BRISTOL	2,537	3.40	2,498,282	6.40
SUNAPEE	2,559	3.41	3,753,515	6.57
FREMONT	2,576	3.41	768,185	5.89
BRENTWOOD	2,590	3.41	771,805	5.89
GILMANTON	2,609	3.42	1,493,586	6.17
ROLLINSFORD	2,645	3.42	835,310	5.92
CHESTER	2,691	3.43	1,013,028	6.01
GREENLAND	2,768	3.44	1,204,935	6.08
NOTTINGHAM	2,939	3.47	1,091,841	6.04
MOULTONBOROUGH	2,956	3.47	3,506,437	6.54
STRAFFORD	2,965	3.47	893,047	5.95
CANAAN	3,045	3.48	1,839,983	6.26
WAKEFIELD	3,057	3.49	1,805,051	6.26
BARNSTEAD	3,100	3.49	1,608,588	6.21
CHESTERFIELD	3,112	3.49	2,286,910	6.36
WILTON	3,122	3.49	2,033,603	6.31
DEERFIELD	3,124	3.49	1,937,144	6.29
NORTHWOOD	3,124	3.49	1,418,169	6.15
GORHAM	3,173	3.50	3,827,227	6.58
NEW LONDON	3,180	3.50	3,462,182	6.54
WALPOLE	3,210	3.51	2,425,109	6.38
NEW BOSTON	3,214	3.51	1,932,824	6.29
TILTON	3,240	3.51	1,719,738	6.24
ALTON	3,286	3.52	2,958,199	6.47
OSSIPEE	3,309	3.52	4,037,042	6.61
NEWTON	3,473	3.54	1,414,819	6.15
LANCASTER	3,522	3.55	5,302,740	6.72
CANDIA	3,557	3.55	1,215,657	6.08
BOSCAWEN	3,586	3.55	1,440,818	6.16
EPSOM	3,591	3.56	1,129,369	6.05
NORTH HAMPTON	3,637	3.56	2,621,011	6.42
MILTON	3,691	3.57	1,640,989	6.22
PITTSFIELD	3,701	3.57	2,221,767	6.35
LEE	3,729	3.57	1,623,753	6.21
HINSDALE	3,936	3.60	1,797,216	6.25
ENFIELD	3,979	3.60	3,301,133	6.52
NEW IPSWICH	4,014	3.60	1,431,500	6.16
WINCHESTER	4,038	3.61	2,301,339	6.36
SANDOWN	4,060	3.61	1,504,893	6.18
AUBURN	4,085	3.61	1,373,330	6.14
LOUDON	4,114	3.61	1,752,196	6.24
HENNIKER	4,151	3.62	2,616,758	6.42
HAVERHILL	4,164	3.62	1,252,612	6.10
NORTHFIELD	4,263	3.63	1,809,819	6.26
HILLSBOROUGH	4,498	3.65	6,107,675	6.79
RYE	4,612	3.66	4,159,532	6.62

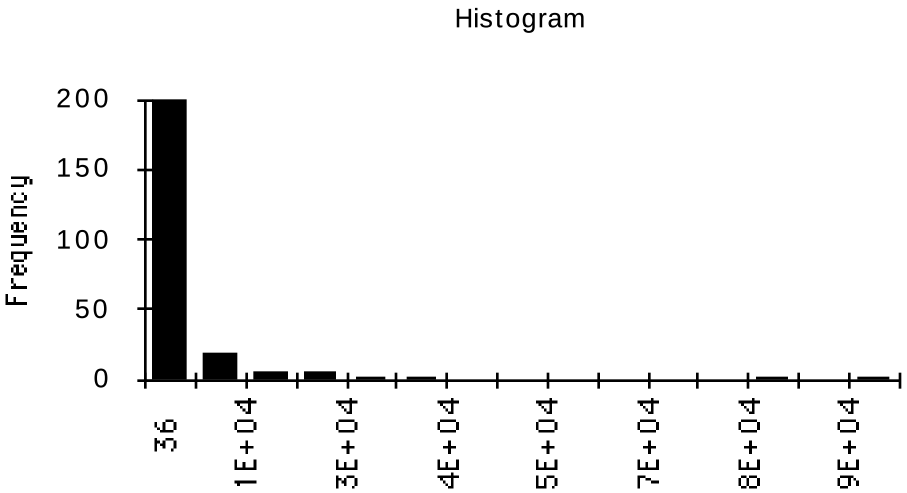
CHARLESTOWN	4,630	3.67	2,602,431	6.42
ALLENSTOWN	4,649	3.67	2,352,666	6.37
HOPKINTON	4,806	3.68	2,856,304	6.46
WOLFEBORO	4,807	3.68	11,206,835	7.05
MEREDITH	4,837	3.68	6,951,966	6.84
RINDGE	4,941	3.69	1,841,645	6.27
STRATHAM	4,955	3.70	2,125,127	6.33
EPPING	5,162	3.71	2,423,933	6.38
ATKINSON	5,188	3.71	2,125,331	6.33
PETERBOROUGH	5,239	3.72	4,995,961	6.70
JAFFREY	5,361	3.73	6,983,140	6.84
BOW	5,500	3.74	4,110,238	6.61
LITCHFIELD	5,516	3.74	1,725,133	6.24
KINGSTON	5,591	3.75	2,254,780	6.35
HOLLIS	5,705	3.76	3,674,216	6.57
FARMINGTON	5,739	3.76	2,610,103	6.42
BELMONT	5,796	3.76	3,352,536	6.53
PLYMOUTH	5,811	3.76	4,317,324	6.64
LITTLETON	5,827	3.77	4,210,515	6.62
GILFORD	5,867	3.77	5,824,931	6.77
NEWPORT	6,110	3.79	5,242,011	6.72
BARRINGTON	6,164	3.79	2,023,182	6.31
WEARE	6,193	3.79	2,865,631	6.46
SWANZEY	6,236	3.79	2,276,072	6.36
SEABROOK	6,503	3.81	12,510,753	7.10
PEMBROKE	6,561	3.82	6,894,823	6.84
HAMPSTEAD	6,732	3.83	2,466,294	6.39
NEWMARKET	7,157	3.85	4,467,355	6.65
PLAISTOW	7,316	3.86	3,451,783	6.54
CONWAY	7,940	3.90	5,988,938	6.78
FRANKLIN	8,304	3.92	6,792,378	6.83
RAYMOND	8,713	3.94	3,804,497	6.58
HOOKSETT	8,767	3.94	7,225,433	6.86
WINDHAM	9,000	3.95	5,286,475	6.72
AMHERST	9,068	3.96	4,672,970	6.67
HANOVER	9,212	3.96	8,928,686	6.95
PELHAM	9,408	3.97	5,087,464	6.71
SOMERSWORTH	11,249	4.05	7,619,638	6.88
MILFORD	11,795	4.07	8,529,696	6.93
DURHAM	11,818	4.07	7,989,037	6.90
BERLIN	11,824	4.07	9,251,705	6.97

LEBANON	12,183	4.09	21,029,363	7.32
HAMPTON	12,278	4.09	13,434,218	7.13
EXETER	12,481	4.10	12,223,190	7.09
BEDFORD	12,563	4.10	9,122,876	6.96
CLAREMONT	13,902	4.14	10,445,650	7.02
GOFFSTOWN	14,621	4.16	8,220,285	6.91
LACONIA	15,743	4.20	11,776,111	7.07
HUDSON	19,530	4.29	13,193,546	7.12
LONDONDERRY	19,781	4.30	13,812,112	7.14
MERRIMACK	22,156	4.35	15,351,214	7.19
KEENE	22,430	4.35	22,915,928	7.36
DOVER	25,042	4.40	20,808,160	7.32
SALEM	25,746	4.41	29,189,769	7.47
PORTSMOUTH	25,925	4.41	20,979,565	7.32
ROCHESTER	26,630	4.43	19,804,052	7.30
DERRY	29,603	4.47	19,299,092	7.29
CONCORD	36,006	4.56	37,787,640	7.58
NASHUA	79,662	4.90	45,170,090	7.65
MANCHESTER	99,567	5.00	87,173,729	7.94
Low Quartile	937	2.97		
2nd Quartile (median)	2,117	3.33		
High Quartile	4,644	3.67		
Median (checking)	2,097			
Mean	4,740			
Sum/234 (Checking)	4,740			

And where does analysis of the relation between population and budget (or the relation between any two variables) begin? With well behaved variables — with an examination of the units of measure, translating if necessary to the forms of well behaved variables.

Well-Behaved Population

First population: Because my spread sheet program is good at putting things in rank order , with names attached, and because it is poor at stem and leaf (and because I value my own time) I will forego the Stem and Leaf and rely on the rank order for equivalent detail. And because my spread sheet program is good at bad histograms and more cumbersome for good ones, I will settle for a less than friendly histogram — it will suffice to give me an overview of the shape of the distribution. Using the population data, as given (with the person as the unit of measure), here, is the histogram. It is not well behaved; on the contrary, it is extremely skewed with a tail extending in the direction of the larger values.

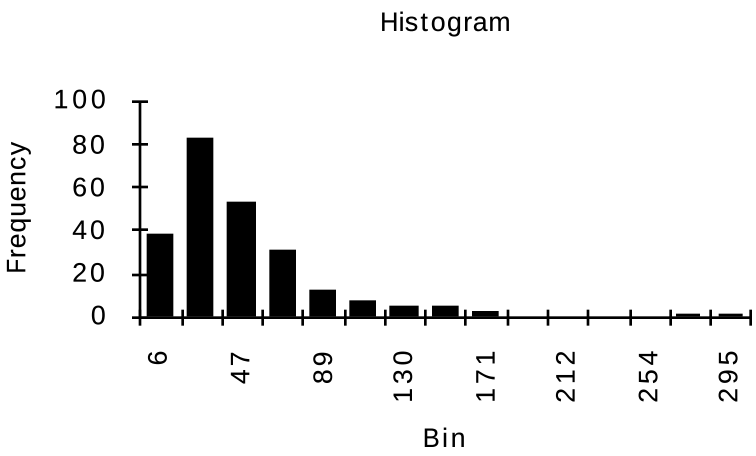


Putting numbers (on what is already obvious from the picture), the mid values give numerical expression to the skew in this picture: The mid-quartile is larger than the median. The mid-eighth is larger than the mid quartile, and so forth.

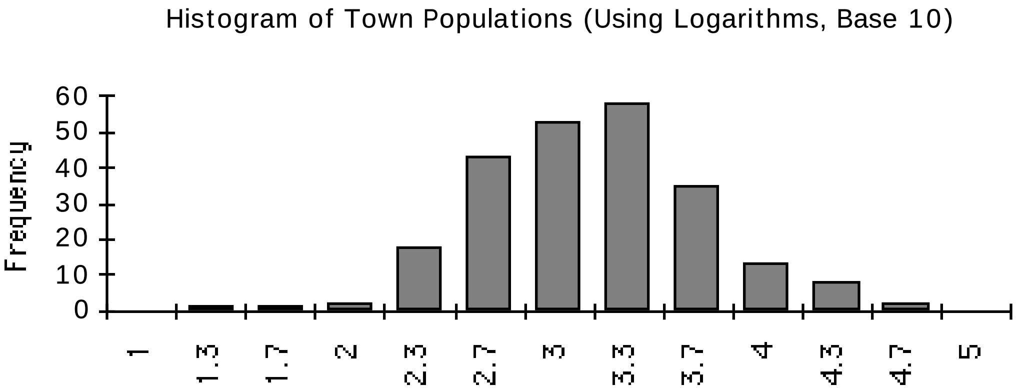
Count						
n=234	Population	Population	Mid Value		Examples	
117.5	2,116.5	2,116.5	2,117	Median	Troy; Sanbornton	
59	935	4,649	2,792	Mid Quartile	Freedom	Allenstown
30	624	8,304	4,464	Mid Eighth	Piermont	Franklin
15.5	356.0	13,232.5	6,794	Mid Sixteenth	Landaff; Easton	Bedford; Claremont
8	248	25,042	12,645	Mid Thirty-Second	Roxbury	Dover
4.5	187	28,103	14,145		Waterville Valley; Easton	Rochester; Derry
2.5	91	57,834	28,962		Ellsworth; Windsor	Concord; Nashua
1	36	99,567	49,802		Harts Location	Manchester

Manchester at 99 thousand people is, by itself, equal to the cumulative populations of the 107 smallest towns. That makes it large, but not necessarily different.

Pursuing the well-behaved form, changing the unit of measure from the person to the square roots improves the symmetry, but not enough.



Pursuing the well behaved form by changing to the logarithm, it is clear (to the eye) that logs are close



Putting numbers on the image, the mid value numbers support the visual appearance. For comparison, the table below shows 3 transformations, including the logs as well as two power transformations, one a little weaker than the logarithm, one a little stronger. Mid values based on the weaker transformation, the .1 power, still show a slightly increasing trend of values, still indicating a tail in the direction of the larger population values. Mid values based on the logarithmic transformation wander — as they will when data are symmetrical. Mid values based on the stronger transformation, the -.1 power, also wander, like the mid values for the logarithm. So the negative .1 power is also close. That makes it user’s choice: I’ll use the logarithm as the well behaved unit of measure for these populations.

power				
0.1	2.15	2.15	2.15	Median
	1.98	2.33	2.15	Mid Quartile
	1.90	2.47	2.18	Mid Eighth
	1.86	2.58	2.22	Mid Sixteenth
	1.74	2.75	2.24	Mid Thirty-Second
	1.68	2.78	2.23	
	1.57	2.97	2.27	
	1.43	3.16	2.30	

log				
	3.33	3.33	3.33	Median
	2.97	3.67	3.32	Mid Quartile
	2.80	3.92	3.36	Mid Eighth
	2.79	4.12	3.46	Mid Sixteenth
	2.39	4.40	3.40	Mid Thirty-Second
	2.26	4.45	3.36	
	1.95	4.73	3.34	
	1.56	5.00	3.28	

power				
-0.1	0.46	0.46	0.46	Median
	0.50	0.43	0.47	Mid Quartile
	0.53	0.41	0.47	Mid Eighth
	0.54	0.39	0.46	Mid Sixteenth
	0.58	0.36	0.47	Mid Thirty-Second

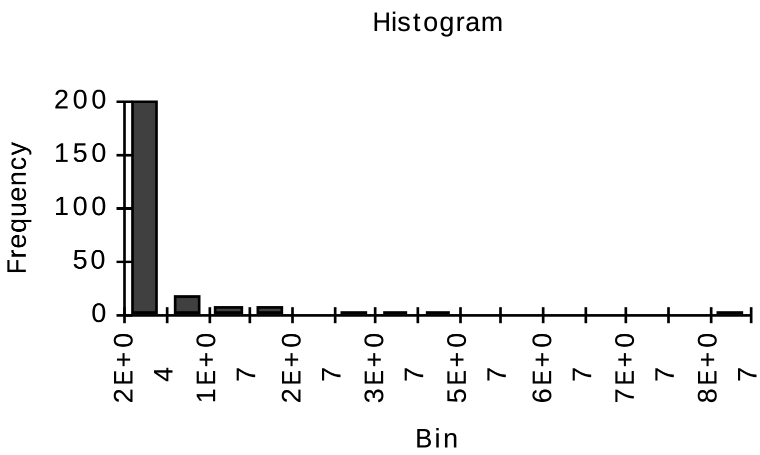
	0.59	0.36	0.48	
	0.64	0.34	0.49	
	0.70	0.32	0.51	

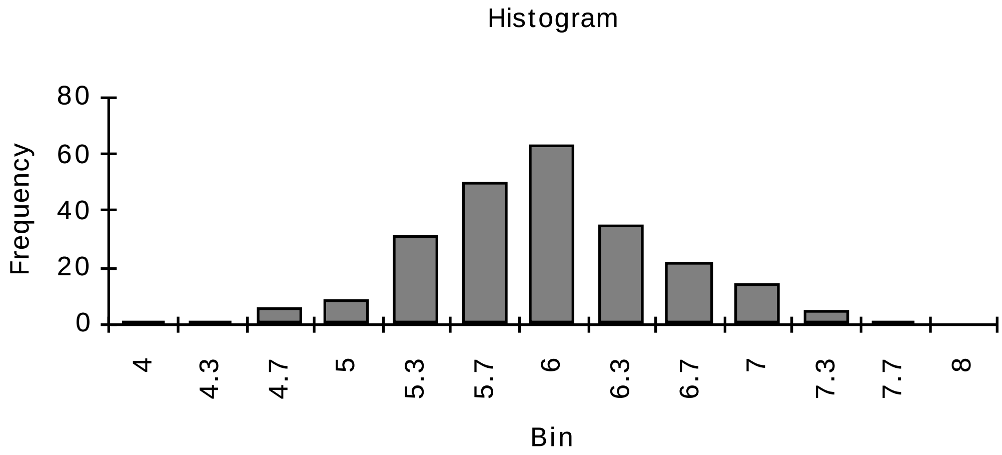
Using logs, and computing the fences, Nashua and Manchester exceed the inner fences on the high end. Nothing is so large that it exceeds the outer fences. On the low end, as on the upper end, two towns are below the inner fences. None are below the outer fences: Using logarithms for the unit of measure, the variable is well-behaved, with a note of caution at each end.

	0.697			Quartile Spread
	1.045			Step Size
In Population	In Logs	In Logs		In Population
84	1.926	4.712	51,544	inner fences
8	0.881	5.757	571,483	Outer Fences

Well-Behaved Appropriations

That is one variable done, one more to go. For Appropriations: Using dollars as the unit of measure, the distributions is, like its mate, sharply skewed, with a few large values at the high end. Using logarithms as the unit of measure the behavior changes, close to symmetry. Attempting to verify this with the mid values, the result, using logs, is disconcerting. The numbers do not support what the eyeball has suspected. the distribution is not really symmetrical: It shows a consistent trend of mid values, 6.10, 6.11, 6.18, etc., indicating that even using logs as the unit of measure, the distribution is skewed with a tail extending in the direction of the higher appropriations.


Histogram
Frequency
200
150
100
50
0
2E+04
1E+07
2E+07
3E+07
5E+07
6E+07
7E+07
8E+07
Bin


Histogram
Frequency
80
60
40
20
0
4
4.3
4.7
5
5.3
5.7
6
6.3
6.7
7
7.3
7.7
8
Bin

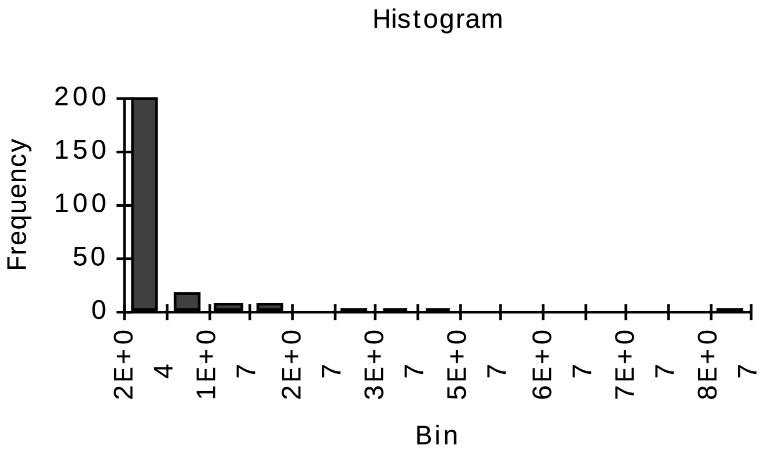
Count						
n=234						

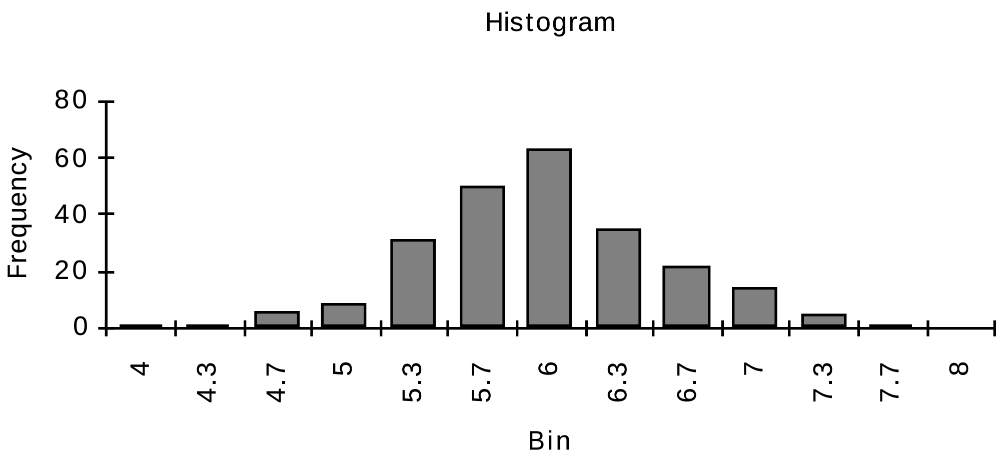
117.5	6.10	6.10	6.10	Median	Haverhill;Plainfield	
59	5.76	6.46	6.11	Mid Quartile	Unity	Newington
30	5.51	6.84	6.18	Mid Eighth	South Hampton	Pembroke
15.5	5.34	7.09	6.22	Mid Sixteenth	Errol;Randolph	Exeter;Seabrook
8	5.14	7.32	6.23	Mid Thirty-Second	Sharon	Dover
4.5	4.82	7.41	6.12		Windsor;Roxbury	Keene;Salem
2.5	4.62	7.80	6.21		Ellsworth;Benton	Nashua;Concord
1	4.20	7.94	6.07		Harts Location	Manchester

Can I accept that as symmetry? Is it close enough? I don’t know. The only way to answer the question is to try a stronger transformation of the unit of measure and see how strong a transformation it takes to eliminate this trend among the mid values.

What it takes to break the trend is the -.11 power. So my choice is between the logarithm and the negative -.11 power. I’ll take the log as close enough. (Had it required the -1 power to break the trend, or even the -.5 power, I would have worried. Fortunately I don’t have to figure out what I would have done.) (Note that the -.11 power reverses the rank order of the numbers, high becomes low and low becomes high. So the “high values” at the mid thirty-second and later correspond to low values on the original scale of the variable. If this were a tail, it would be a tail toward the small values appropriations, implying that the transformation had been too strong.)

power						
-0.11	0.213	0.213	0.213	Median		
	0.232	0.195	0.214	Mid Quartile		
	0.248	0.177	0.212	Mid Eighth		
	0.259	0.166	0.212	Mid Sixteenth		
	0.272	0.157	0.214	Mid Thirty-Second		
	0.289	0.153	0.221			
	0.311	0.145	0.228			
	0.345	0.144	0.244			





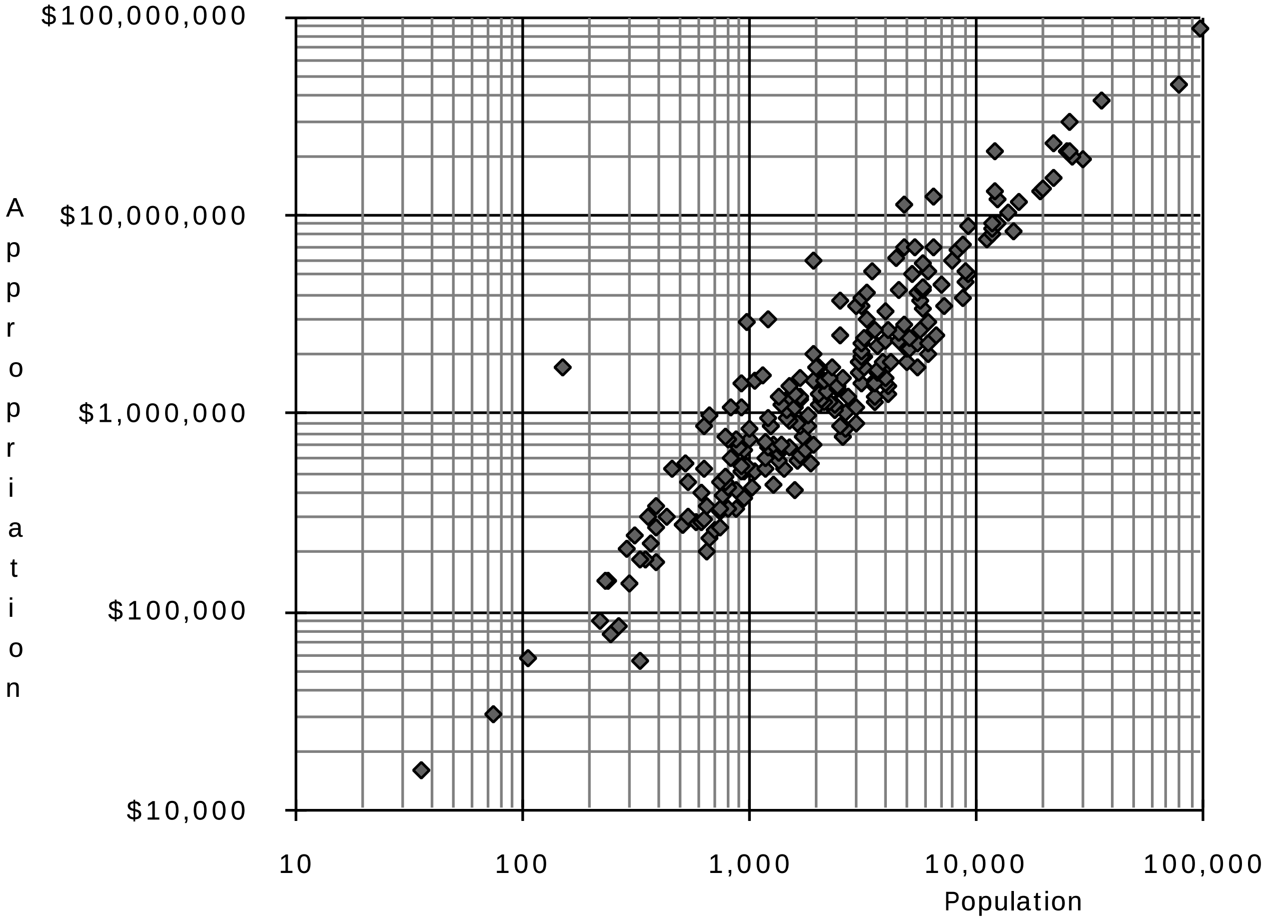
The Relation Between Population and Total Appropriation

Is there a relation between population and appropriation? That was the question I had to answer. If there is a relation? And if the relation is a relation of strict proportionality, double the population and you will double the appropriation (on the average) — then I have support for the procedures of mean-based budgeting: Strict proportionality, empirically, establishes reason to use appropriations per capita as a standard for the budget — regardless of the size of the town.

Searching each separate variable for its well-behaved form, tells me to look for the relation between these two variables by examining their log log form, using the log form of each variable.

Ordinarily, I would actually do the log log graph on "wallpaper", making it about three feet wide. It takes a graph of this size to provide proper labels that literally spell out the name of each town on the graph Then I would feast my eyes on the result, locating Waterville Valley, a ski town in the mountains, locating Manchester, an old industrial town, locating Concord, the state capitol, Hanover, a college town, ... getting a "feel for the data".

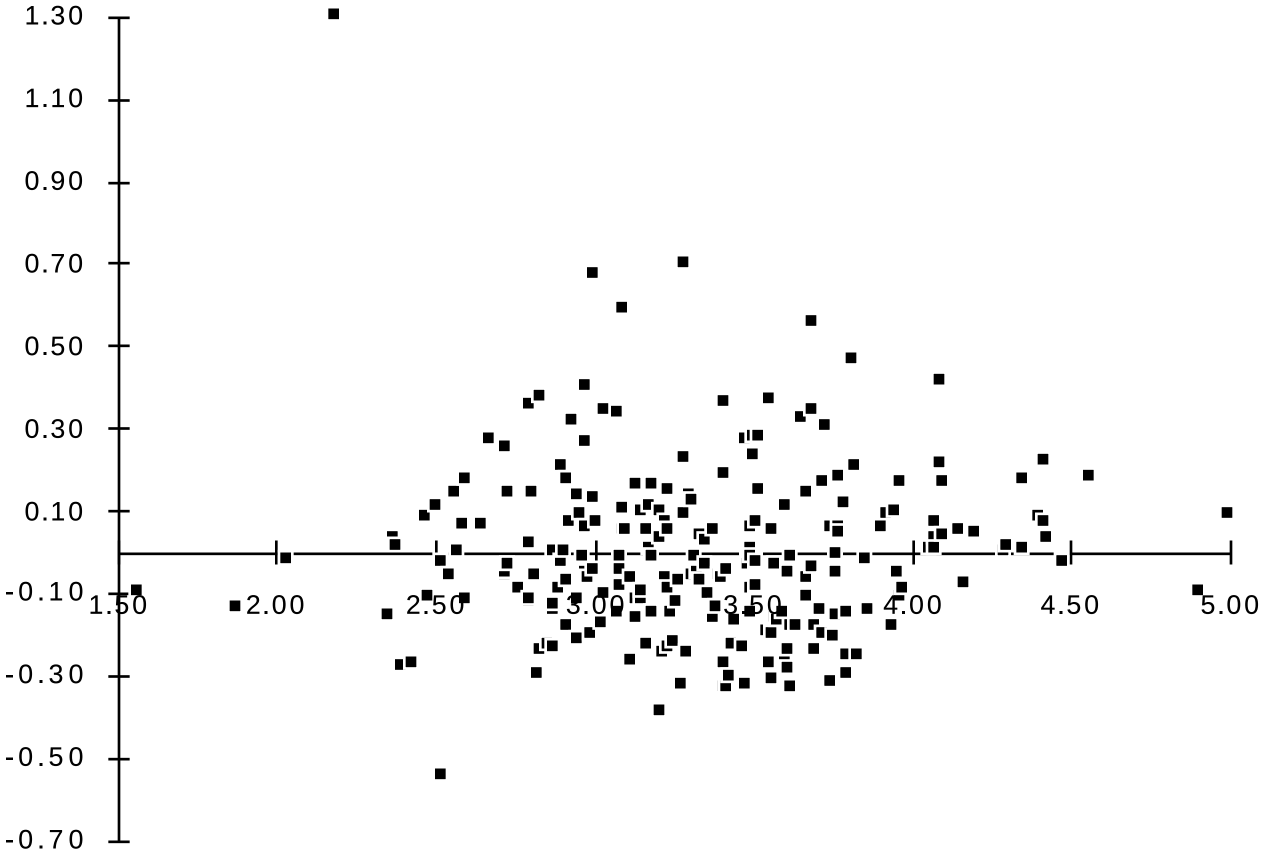
But ,limiting myself to a publishable piece of paper, I can at least inspect the shape of the graph, using dots.



My first reaction to the graph is relief: So far , without getting serious (that is, without looking at the residuals) it appears that the relation is a line, not a curve. And the slope is tantalizingly close to 1, at a height showing that a population of 1,000 will have, on the average an appropriation of about $600,000, about $600 dollars per person. There is hope: This could be a linear relation, in logs, and a strictly proportional relation in dollars and people.

Now I've got to get serious, using least squares regression to estimate a line at the center of the cloud of data and then looking at the residuals. So, allowing my spread sheet program to execute a least squares regression of log appropriations on log population, using common logs (base 10), I get an estimate of a line at the center of this cloud of data, a line with intercept 2.689 and slope of 1.032.

That is a little worrisome: I note that the anti log of 2.689, base 10, is 489, $489 per person — a little low compared to my eyeball estimate of $600. But then, checking, $600 per person would have corresponded to an intercept of 2.778. The difference between 2.689 and 2.778 may be too fine for eyeball discrimination. So $4889 may be acceptably close to my original estimate. Looking at the slope, the slope at 1.032 is also nice — passably close to a very simple number, passably close to the magic number 1 — at which I can establish strict proportionality. Using this intercept and this slope, as estimated by the computer, and plotting residuals (represented in logs), the residuals are



The plot of the residuals makes the extremes in appropriations per capita stand out. I see that exceptional point at the top of the graph and I have to look: That "dot" is Waterville Valley, site of a large ski development in the White Mountains.

		In logs	In Ratios

Low Quartile	2.97	-0.14	0.72
Mean	3.33	0.00	1
High Quartile	3.67	0.10	1.25
		In logs	In Ratios
	Quartile Spread	0.24	1.73
	Step Size	0.36	2.28
	lower inner fence	-0.50	0.32
	Lower Outer Fence	-0.86	0.14

The mean residual is 0, as it must be in least squares regression. In logs, using the quartiles, fifty percent of the towns lie in a range between -.14 and +.10. In ratios those two numbers translate into a ratio that is less than one, .72, and a ratio that is greater than one, 1.25. In percentages, that means that fifty percent of the town budgets like in a range from 28% below the average to 25% above the average of appropriations per capita for the entire state. Looking for outliers, one of these residuals is below the inner fence on the low end. But, as is obvious on the graph, there are several outliers exceeding the fences at the high end of appropriations per capita — five exceed the inner fence of appropriations per capita, one of the five exceeds the outer fence: That is Waterville Valley again, with 151 in its official population, and budget of $1.7 million dollars — that is $11,500 per capita, definitely an outlier.

Leaving this one point out, and re-estimating, the revised estimates are now intercept = 2.604 ($402 per capita), slope = 1.056.

Interpreting the slope itself, in logs the relation is

log (Appropriation) = 2.604 + 1.056(log population)

Taking anti-logs on both sides of the equation and restoring the original units, that is

Appropriation = ($402per person)x(population$^{1.056}$)

Thinking about the Exponent: The Relation Between Population and Budget

Thinking: How would I *like* this relation to turn out? That's clear. I want the exponent to be one. That would establish that the relation between budget and population is independent of the size of the city. That would gives me empirical support for the standard by which I can begin to sort the budgets of the state of New Hampshire and, in particular, the budget of my town.

Carl Sagan suggests that the essence of scientific method is skepticism. I am surely a skeptic — I don't even trust myself. That's why it is important to be up front about what I would like to find and, therefore, to be particularly careful and suspicious when, low and behold I find, at the end of my analysis that I have "discovered", exactly what I was looking for at the beginning. Maybe, but I have to be careful.

So what I have so far is a proportionality between Appropriations and the 1.056th power of population, not the first power. Can I just lop off that .056, declare the power to be 1 (close enough). If I were to do that I would argue something sophisticated like

> "On grounds of parsimony, I will simplify that 1.056 to 1, which establishes that the appropriations are directly proportional to the population."

But how do I know that .056 is small? Compared to what? Like analyzing budgets, data analysis is often a “game” of establishing contexts. How do I know that .056 is small? Can I leave it out?

I will figure that out by trusting the mathematics and getting a feel for it. What 1.056 says, in contrast to 1.000, is that larger cities spend more, per capita (and on the average), than small towns. How much more? Following the math, I will figure out the numbers with and without that .056. Using strict proportionality (using 1.000), suppose that a town of one thousand people could expected to appropriate $400,000. Using strict proportionality, by comparison, a town of 10,000 people would be expected to appropriate ten times more, $4,000,000, and a town of 100,000 (Manchester) would be expected to appropriate one hundred times more $40,000,000.

Log Population	Population	Prediction in logs	Prediction in dollars	Ratio (to first row)	Ratio of the appropriation to the appropriation of a town of 1,000 people.
3	1,000	log appropriation = intercept + log population	$appropriation $= 10^{intercept} populatio$		1
4	10,000	log appropriation = intercept + log population	$appropriation $= 10^{intercept} populatio$	$\frac{10^{intercept} 10{,}000}{10^{intercept} 1{,}000} = 10$	10 to 1

5	100,000	log appropriation = intercept + log population	$appropriation $= 10^{intercept} populatio$	$\frac{10^{intercept}100{,}000}{10^{intercept}1{,}000} = 100$	100 to 1

Now, by contrast how big is that 1.056? ".056" looks small but, actually the increase is not obviously so small that it can be ignored: Compared to a town of 1,000 people, a town that is 10 times larger would have an appropriation that is larger by the ratio of (10^1.056)/(1^1.056). That is 11.38 to 1. And it means that the relative budget is 14% larger than would be expected under strict proportionality. And Manchester, the extreme at with approximately 100,000 people would have an expected budget that is 29% larger than would be expected under strict proportionality.

Log Population	Population	Prediction in logs	Prediction in dollars	Ratio (to first row)	Ratio of the appropriation to the appropriation of a town of 1,000 people.
3	1,000	log appropriation = intercept + 1.056 log population	$appropriation $= 10^{intercept} population^{1.056}$		1

4	10,000	log appropriation = intercept + 1.056 log population	\$appropriation $= 10^{intercept} population^{1.056}$	$\frac{10^{intercept}10{,}000^{1.056}}{10^{intercept}1{,}000^{1.056}} = 10^{1.056}$	11.38 to 1
5	100,000	log appropriation = intercept + 1.056 log population	\$appropriation $= 10^{intercept} population^{1.056}$	$\frac{10^{intercept}100{,}000^{1.056}}{10^{intercept}1{,}000^{1.056}} = 100^{1.056}$	129.4 to 1

So, I can't casually throw it away: With or without that .056 tacked on to the 1 (in the exponent), I would not or would find a 29% larger budget (to be typical or large). It is not a large amount. It affects only one large city and (at this magnitude) it only affects the comparison between the largest city and the smaller towns, not the comparison to the average. But it is worth attention.

Now I'm going to tackle it another way, asking "How much do I believe these numbers anyway?" These numbers are the facts, not a sample of the facts, so variability is not an issue. But I note that just removing one data point raised the slope from about 1.04 to about 1.05. I don't really believe that number out to as many digits as I can calculate. Is the contrast between 1 and a slope of 1.05, on the log log graph, within the "wobble" or uncertainty that is built in to my data? The slope and the intercept estimated by my computer minimize the squared deviations. How much larger would the squared deviations become if I were to impose a slope of 1.000? How sensitive is the squared error (by which least squares regression evaluates the result) to the contrast between the simple 1 and the observed slope of 1.056?

Starting with the best, starting with 1.056, the equation “explains” 89.07% of the variance:

By regression, r = .9438, r2 = .8907, 89.07 percent of the variance is “explained”.

Intercept = 2.604,
slope = 1.056.

By contrast, imposing the exponent 1.000, the equation “explains” 88.77% of the variance:

Intercept = 2.304 (anti-log of 2.304 = $637),
slope ≡ 1.000

That almost convinces me that the difference can be ignored, 89.1% for the statistical best answer, 88.8% for the simple answer, 1. In fact, “1” may be the better answer. In the world of science I am free to and encouraged to look at details that may be a hint of some subtlety — but in my tentative conclusions I am obliged to choose what is simple unless there is compelling reason to do otherwise. I also note that imposing the slope 1 leads to a re-estimate of the constant so that it now corresponds to $637. That is closer to what I saw in the graph and close to the median $598 per person, which is less subject to error due to single cases like Waterville Valley or the other four high values of the residuals. In a sense by “helping” the statistics with their estimate of the slope, I may have been rewarded with a better estimate of the intercept.

I am not wholly happy, but I am willing to commit:

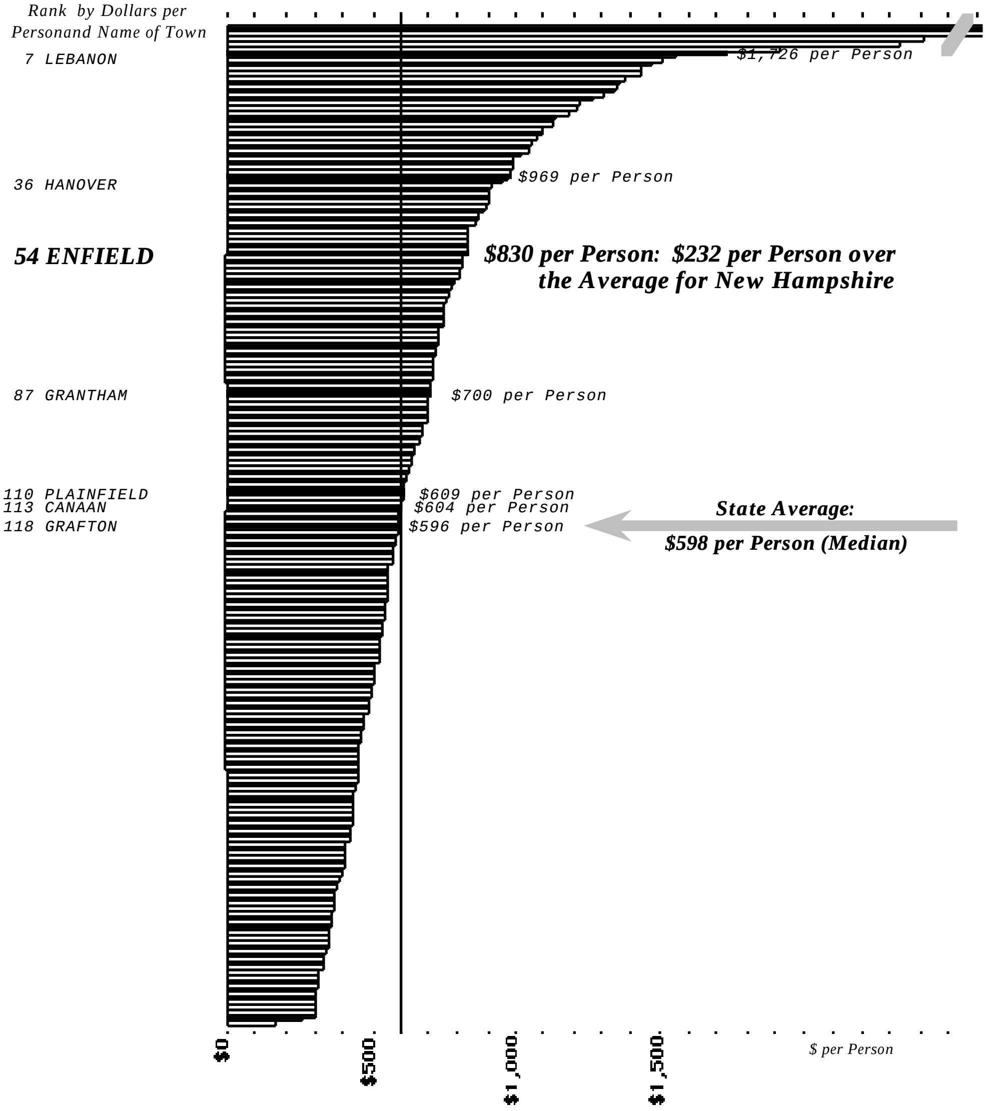
> “On grounds of parsimony, I will simplify that 1.056 to 1, which establishes that the appropriations are directly proportional to the population —at a trivial loss of approximately 1% in the variance explained, and at a gain of considerable simplicity.

I will keep that 1.056 in mind and, in a professional publication I would be sure to alert the reader. But I would not use it unless subsequent research advanced the case for using the more complicated rather than he more precise answer in cases where both “explain” approximately the same amount of the variance.

And here is my first report:

Enfield Town Budget, 1994

Total Appropriation Shown as Dollars per Person, Compared to State and to Adjacent Towns



The Enfield budget is $3,301,133 for a 1990 population of 3,979. Reduced to the state average of $598 per person, the Enfield budget would be to $2,379,442. It would be $921,691 below the present budget.

Appropriation Data from State of New Hampshire Department of Revenue Administration

Prepared by Joel H. Levine, RR1 Box 116, Enfield, New Hampshire 03748 8/1/95 b

The Ego of the Data Analyst — Postscript

I note, with pain, that this report use few words at all. My intent in this case was to build a chart, one page, few words, that would begin a discussion. I know, and in a report written to professionals you would know, that the comparisons implied by the chart are valid. I know and you know that it is valid to standardize budgets by computing the appropriation *per* person and to compare them in that form. And if someone had chosen to take up the question of validity, I would have been ready. It is not easy work or, to put it another way, you have to be up to a certain level of competence before it *is* easy work. But for the most part, few people will care. On the other hand, if someone does begin to ask you the right questions, you and the interrogator will, each of you, have found a worthwhile colleague.

Try it. Data grouped by the broad classifications of the state accounting categories are enclosed. Without ever setting foot in the Town of Enfield, without hours spent over the budget and discussions with the accountants, you can easily show that there was a $400,000 purchase in need of explanation. Without any attention to the local press you will find that someone really should say a few words about a $6,000,000 piece of goods purchased by the Town of Lebanon. These are simple statistical outliers brought into focus by the application of mean based budgeting.

____

3

Calculus and Correlation: A Calculus for the Social Sciences

Thirty years ago optimistic social scientists dreamed of a calculus of the social sciences. The logic of this calculus would mirror the logic of social events and it would be the natural language with which to formulate laws of behavior. With it, social science would become "science." That distant dream expressed both admiration for physical science and hope. The admiration was for the incredible fit between the calculus and basic concepts of physical science, a fit so apt that it is no longer possible to think about phenomena like "velocity" and "acceleration" apart from their mathematical expressions. And, seeing that fit, social scientists hoped for a calculus of our own, a mathematics that would confer upon us the blessings that classical calculus had bestowed upon physics. Our present generation of survey researchers would be our Tycho Brahes, documenting the basic facts of social existence. Our Newton would lie just ahead, discovering law.

Why invent a calculus of our own? Why not borrow? Presumably because social science is built from variables like class and power, from roles like "mother," from religion and politics, and because such variables defy the scientific heritage of physics. The science of these things will require a new math. That's one view. But one way that science moves forward is by looking backward to re-assemble its understanding of what it has already accomplished and I'm going to look backward to argue that the calculus, or a calculus, of the social sciences is already here — even for such variables. What's more, I suggest that the Newton and

Leibniz of our calculus are none other than the original Isaac Newton and Baron von Leibniz — because it is the same calculus.

My theme is that our basic idea of correlation, in the social sciences, and their basic idea of the derivative, in the calculus, are fundamentally the same. The similarity is well disguised by convention and context, but the two are essentially the same. In one sense my task is simple: I have to show the correspondence between the two forms of the one concept. In another sense the task is difficult because I have to break through the disguises of convention and context, playing with uncomfortably basic questions like what *is* correlation and what *is* a derivative. It's tricky to deal with broad generalities while retaining useful content, but that's the task. The immediate implications are for the concept of correlation: Currently, two-variable correlation is reasonably well understood among social scientists. But partial correlation is something we know more by practice than by intellect, and three-variable, four-variable and n-variable correlation are hardly understood at all, at present. James A. Davis put it well in successive chapter headings for his book on survey analysis, referring to “Two Variables,” “Three Variables,” and “Too-Many Variables.”[1] But we can do better than that. If we rely on the ideas of the calculus as the organizing principle, then the whole intellectual picture of correlation becomes “obvious.”

The Logic of Contrasts: Change of One Variable

Let's begin on neutral turf, neither correlation nor the derivative, beginning with the basic problem of comparing two numbers. Suppose that the great oracle who prepares data for data analysts encodes a sequence of four messages about the world, shown in Figure 3.1.

My job as analyst is to read these four messages and interpret them. But, oracles being what they are, the messages are both badly labeled and ambiguous, leaving me little to work with. One fact that's clear about them is that the two numbers in each pair are different: 2 is not

[1]. James A. Davis, *Elementary Survey Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1971.

		Reading	Reading
Message	Message	One:Two:	
		Ratios	Differences
#1:	[2, 4]	2	2
#2:	[4, 8]	2	4
#3:	[6, 12]	2	6
#4:	[8, 16]	2	8

Figure 3.1
Four Sets of Data Read as Messages Describing the Real World

equal to 4; 4 is not equal to 8. So I can ask the question, “How different are they?” If I compare the two numbers by their ratio, then I think the great oracle is telling me “2” in the first message, and then “2” again, and again “2.” That's clear enough. Reading the messages as fractions, the sequence of four messages is constant. The data vary, but the message is constant.

Unfortunately, there is another obvious interpretation for the same data and if I compare the two numbers by their difference, instead of their ratio, then I get a conflicting result: If I compare two numbers by their difference, then I think that the great oracle is telling me “2” and then “4,” “6,” and “8.” That's also clear. Reading the messages as differences, the sequence is a regular progression, not a constant. And, of course, having found two ways to read the data I can probably find more, and that leaves me with a choice: Which is correct? Do the messages describe a constant, or a progression, or something else? Which reading of the data connects to reality? Without labels and clarity, the question is undecidable and that's the point: The numbers do not speak for themselves. Here at the very lowest level of quantification, comparing numbers, they require human intervention and choice. Quantification is about ideas. Get the numbers right with respect to reality and you have a foundation that lets you progress to the next level of difficulty. Get them wrong here, at the beginning, and any subsequent analysis of the data, and of the reality behind the data, is in jeopardy.

$$\lim_{x \to x_o} \frac{f(x) - f(x_o)}{x - x_o}$$

Figure 3.2
The Derivative of "*f*" with Respect to "*x*"
Expressed as the Ratio of Two Differences

But in one important sense there really is a universal answer to the question, or at least a preferred answer, for all cases: How different are two values of one variable? I'll have to justify this answer, but as a matter of form *the correct answer is always their difference*: Never their ratio, never a percentage, always the difference. Always, for any variable and any problem. And the way you make this rule work is by including a generous escape clause: If theory or intuition suggests ratios, then use logarithms before computing the differences of the logs. Whatever it takes, logarithms, inverses, whatever: Transform the variable and then compare values by computing their difference.

Such abstraction may seem distant from the work of a practical scientist — discussing universal answers and content-free rules. But there is solid practical reason for it. This is the kind of thing that mathematicians learn, but rarely teach, because the message is only implicit in the mathematics. It's implicit, for example, in the derivative — the mathematical form, content free, by which we describe velocities and accelerations of all kinds of particular cases. Take a look at a derivative, defined in Figure 3.2, and note the abstraction. The derivative uses two variables, an "*x*" and an "*f*," and each variable has two values. In the denominator, the derivative compares two values of *x*, using subtraction. In the numerator, the derivative compares two values of *f*, again using subtraction. Thus, if x_0 is the starting time of an experiment and *x* is the current time, then the denominator, $x - x_0$, is the time elapsed. And if $f(x_0)$ is a location at time x_0 and $f(x)$ is a location at time *x*, then the numerator, $f(x_0) - f(x)$, is the distance between locations. Putting the two variables together, the derivative divides one difference by the other, giving the rate of change in units of distance divided by time, in miles per hour or centimeters per second.

Part of the genius of the derivative lies in the stunning simplicity of this rate of change at the heart of the concept. A rate of change expressed as "miles per hour" is so obvious to us now, hundreds (or thousands) of years after its invention — so built in to the culture, that math students go right past it, focusing on the less familiar concept of limit, also built in to the definition. But there was a time, I'm told, when physical scientists weren't at all sure that a precise concept of velocity was important. If, for example, you hadn't observed that falling objects fall at different velocities — if you thought they all did the same thing under the influence of what we now call gravity — then you might not be well focused on the concept of velocity. It probably took some serious thinking to define "velocity" as a rate of change and serious creativity to re-use the same idea for acceleration, defining it as the rate of change *of* a rate of change.

What's so clever about the derivative and the rate of change? What's clever is the foresight to begin with subtraction. The derivative compares two values of one variable by subtraction. You would think that flexibility might be a virtue here, when it comes to comparing numbers: When social scientists compare two people's educations, or two counts, or two prices of a commodity, we are ready to use differences, or ratios, or percentages. Whatever makes sense, we'll use it. Compared to our flexibility, in the social sciences, Newton and Leibniz's derivative is the work of monomaniacs, monomaniacs with foresight.

You have to violate the rule of subtraction, temporarily, in order to demonstrate the trouble that would develop without it. For example, let me attempt the mathematical description of a population that doubles each year, and let me compare population to population by division instead of subtraction. Having specified, at the outset, that the population doubles each year, you would think that division would be the natural arithmetic for this population. But watch: If I compare population to population using division, then I have to compare change of population to change of time with an exponent. Thus, if I begin the comparison with division, I have to follow up with the unpleasant expression at the left of Figure 3.3.

This expression does the job, but one thing that Newton and Leibniz did *not* do is work with this kind of thing. It's not just "ugly." The problem is that to generalize this expression, to create a calculus that could calculate its implications would require a whole new branch of the calculus, just for this expression. And mathematicians and physicists

$$\left(\frac{f(x)}{f(x_o)}\right)^{\frac{1}{(x-x_o)}} = 2 \qquad \frac{\log f(x) - \log f(x_o)}{x - x_o} = \log 2$$

Using Division Using Subtraction

"Population doubles each year" — expressed by division, left, and by subtraction, right, where "*f*" is Population and "*x*" is time.

Figure 3.3
Doubling of Population

$$\left(\left(\frac{f(x)}{f(x_o)}\right) \Big/ \left(\frac{x}{x_o}\right)^{\text{Constant}}\right) = 1 \qquad \frac{\log f(x) - \log f(x_o)}{\log x - \log x_o} = \text{Constant}$$

Using Division Using Subtraction

"Proportional growth of the population is directly proportional to the growth of the food supply" — expressed by division, left, and by subtraction, right, where "*f*" is population and "*x*" is food supply.

Figure 3.4
Proportional Growth of Population and Food Supply

don't do that, not if they can avoid it. They keep it simple, which means, in this case, being rigidly simple-minded with respect to form, by subtracting. How do I replace division by subtraction? Using logs. That is what logarithms do for positive numbers, replacing multiplication and division of the original numbers by addition and subtraction of their logarithms. If I transform population to log population and subtract, then the whole concept reverts to standard form as a rate of change, Figure 3.3, where it is subject to well-known methods of inference and easily handled by anyone with a term of calculus.[2]

And if this example doesn't worry you, then try something more troublesome: Try expressing something like “the rate of growth of a population is directly proportional to the rate of growth of the food supply.” If I compare population to population and food supply to food supply using division, then the comparison between change of population and change of food supply requires the distinctly unpleasant expression on the left of Figure 3.4, another new object requiring another new mathematical development. By comparison, changing the numbers to their logarithms and subtracting re-expresses the same concept in standard form as a rate of change. Keeping it simple, the derivative conserves the user's intellectual muscle for something more productive, further down the road.

The popular culture views math as something complicated, but the truth is that math places high value on simplicity, acknowledging the humble fact that human beings have limited cognitive capacity. And the trick to doing better is not to work yourself into a frenzy to make yourself a smarter human being. The trick is to work on the problem to make it simpler — without any change in the problem. Mathematics keeps it simple.

To be sure, there is a price to be paid for simplicity: You have to switch from population to log population, from the kind of thing you can count, one person, two people, three . . . to a different kind of unit. That

[2]. Working with the calculus, on the left, the math establishes the mutual implication between this rate of change and exponential growth. Working with the thing on the right, the same implication is there — it must be. But it is difficult to prove.

isn't easy. Everyone knows what a person is. But a log-person? Everyone knows what a dollar is. But a log-dollar? That's uncomfortable. It takes time to get used to such things. But it pays off with simplicity: Even if you have no direct interest in equations and take no joy from the beauty of the math, even if your interests are immediate and practical, change the variables. Suppose, for example, that I am watching the price of a corporation's stock on Wall Street — that's as immediate and practical as I can get: Suppose I buy a security priced at $100 and observe a change of price to $110. Everyone in this culture knows about percentages. So how could the difference of logarithms, which I recommend, be simpler than percentages as a description of this change? The answer is that percentages are not simple, just familiar, and if you insist on using them for practical work you quickly get into as much trouble as you would if you insisted on using them in equations. Suppose I buy this security at $100 and observe its changes, using percentages. And suppose its price moves up ten percent one day, down ten percent the next, up ten percent the next day, and down ten percent again, regularly. That's steady and sounds stable, but if I keep that kind of steadiness going, day after day, in percentages, I will lose about seventy percent of my money in the course of a year: Check the numbers. Ten percent up from $100 is $110. Ten percent down from $110 is $99. Ten percent up from $99 is $108.90. Follow that out for a year and what's left is about $30.[3] In fact, to stay even, expressing change in percentages, I would have to climb ten percent while holding my loss to about nine and one-tenth percent: 10% up and 9.1% down, just to stay even.

If that surprises you, then I've made my point that percentages are not simple. A little thought will show you the problem — the base from which you must compute the percentages keeps shifting. And, seeing

[3]. Assuming 250 days of trading, on the stock exchange, each year, and using as many digits of accuracy as my computer will allow, the sequence begins

100.0000, 110.0000, 99.0000, 108.9000, 98.0100, 107.8110, 97.0299, 106.7329, . . .

and concludes

. . . , 29.6387, 32.6025, 29.3423, 32.2765, 29.0488, 31.9537, 28.7584, 31.6342, 28.4708.

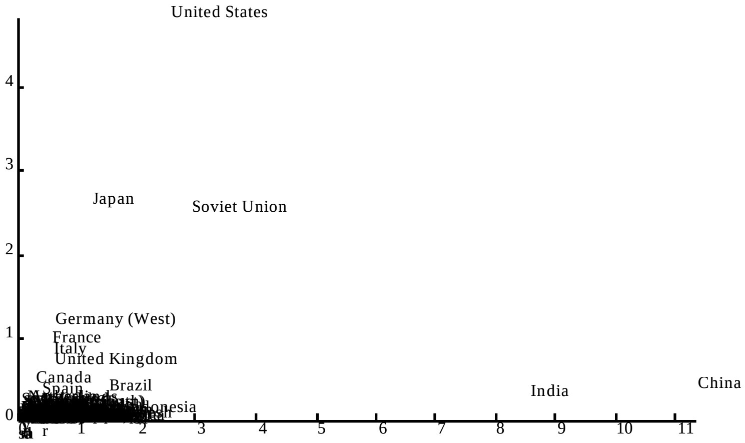
the problem, you might figure out a way to compensate for it. But now, adding some sort of built-in compensation to the discussion, you are no longer talking about a sequence that is simple, using percentages. (In fact, you are probably on your way to re-inventing logarithms.) Better to convert the price to logs and observe, plus log(1.1), minus log(1.1), plus log(1.1), minus log(1.1). Logs are simpler.

Again, I'll admit there is a price: Tell someone, “Today the logarithm of the price of a share of General Motors stock moved up by log(1.1),” and they will look at you strangely. But there is no easy way out. The same people who object to logarithms and claim to be at ease with percentages will look at you strangely when they find that a precise balance of gains and losses, 10% gained, 10% lost, can wipe out 70% of their money in the course of a year.

To follow the same point, that changing the numbers pays off, even in empirical work, consider the data for the relation between the total populations of nations and their gross national products (GNPs). There should be nothing exciting about population and gross national product: Large nations, like China, the former Soviet Union, India and the U.S. should show large populations and large gross national products. (I'm using GNP, not GNP *per capita*.) Using “people” for population and “dollars” for GNP, think about the graph of these two variables. You might expect something like a line extending from low to high, from low population and low GNP to high population and high GNP, from Tokelau (an island nation in the South Pacific) on one end to China on the other. And when you have a clear mental image of the graph you expect for these data, look at the facts graphed in Figure 3.5. In fact, the real graph of “people” (left to right) by “dollars” (bottom to top) is, in non-technical terms, a mess: More than 90% of the nations clump together in a blur at the lower left-hand corner of the graph. The lay explanation for this mess is practical. Practically, it appears that the U.S. and a few large nations are too big, so big that these super-states are different in kind from other nations and don't belong on the same graph. You don’t mix “apples” and “oranges” and you don’t mix Tokelau with China, for which differences of degree have become differences in kind, as one is almost one *million* times larger than the other.

So, following this explanation, you solve the problem by removing the U.S. and other very large nations from the graph and drawing it again, without them. Sounds good, but it doesn't work: If I oblige by removing the special cases, I get the revised graph in Figure 3.6, which is basically a re-run of the same phenomenon, minus a few of the original data points. And if you fix it again, simply by removing more cases, then you will be disappointed, again and again. Holding on to the familiar units of dollars and people, these data will not behave. And each time you compensate for your difficulties, keeping track of what has not been placed on the graph, as well as trying to make sense of what has, the analysis gets more complicated.

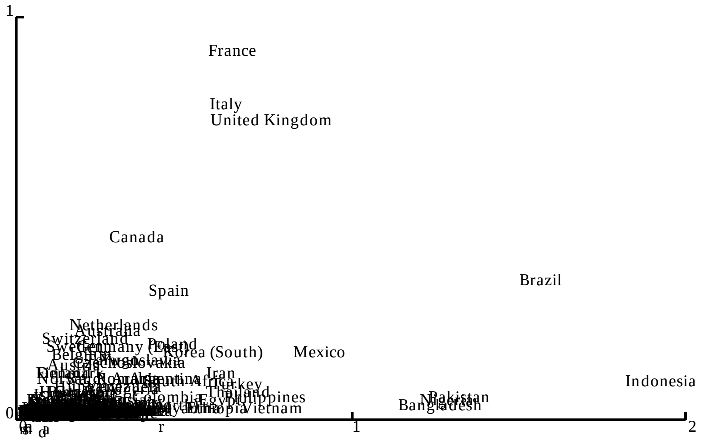
The real problem here is the mathematics: You may not be inter-



1988 population in hundreds of millions, left to right, by 1990 gross national product in trillions of dollars, bottom to top. For example, the United States is plotted at 251 million people and 4.86 trillion dollars. Data from 1991 *Britannica Book of the Year*, Encyclopaedia Britannica, Inc., Chicago, 1991.

Figure 3.5
Gross National Product Versus Population

ested in it — your interests may be practical and data oriented, but you can’t escape it. The practical problem with these data is directly comparable to the mathematical problem in the equations of Figure 3.4. And, fortunately, the solution is also the same: Fix up the comparison between one population and another, and the comparison between one GNP and another, by converting to logs. Then, comparing differences of population to differences of GNP, you will be rewarded with the graph in Figure 3.7. In “theory” the two graphs, Figures 3.5 and 3.7, display exactly the same information. In practice there is a big difference. Even Tokelau, with its two thousand people, and China, with its one billion, line up. Mathematics keeps it simple.



1988 population in hundreds of millions, left to right, by 1990 Gross National Product in trillions of dollars, bottom to top. For example, France is plotted at 57 million people and 898 billion dollars. Data match Figure 3.5.

Figure 3.6
Gross National Product Versus Population — Truncated

First page of Two-Page Spread

Replace with separately printed figure. Print as two facing pages   Figure 3.7

There are two versions of the paste-in pages. They are slightly different at the bottom and require varying tricks in the reduction.

Second page of Two-Page Spread

Replace with separately printed figure. Print as two facing pages Figure 3.7

The Logic of Correlation

In summary, regardless of context and meaning, when you compare two values of one variable, subtract. Change the numbers, if necessary, but subtract. And that simple lesson in the virtue of good form is half of what I need to connect correlation to the derivative. The other half is a good look at correlation, looking at what correlation *is* in the abstract. To anticipate the argument: In the abstract, correlation expresses a particular kind of comparison, involving two or more variables. By custom, however, correlation and its implied comparison, are not generally expressed by subtraction. And that is where the two halves of the argument get joined: When correlation is re-expressed, using subtraction, the form of the comparison shows that both concepts, both correlation and the derivative, are expressing with the same idea.

Let's look at correlation. To illustrate, let me capture the first two messages from my oracle and give them labels. Shown in Figure 3.8, the first message is now a message about “Democrats”: Four of them voted for Bush, two for Dukakis. And the second message is a message about “Republicans”: Eight of them voted for Bush, four for Dukakis. In these hypothetical data, is there a correlation between party and vote? How do you answer the question? Paraphrasing the answer by G. Udny Yule, circa 1903, the answer begins by asking the meaning of a single number:[4] Look at the “2” at the upper left of Figure 3.8, two Democrats voted for Dukakis. Does this single number, “2,” say anything about the correlation between Vote and Party? No, “2” is neither large nor small, except in comparison to some other number.

All right then, reading the first message, Yule would have compared 2 to 4, using the odds: These Democrats voted for Bush, two to one. Now, does this comparison, using odds, say anything about the correlation between party and vote? Still, the answer is no. The 4 to 2 odds, favoring Bush, are neither large nor small except by comparison to the odds among Republicans. The fact that these Democrats favored

[4]. Over the years, Yule's Q, which I am about to derive, has been derived by alternative methods and has been related to other measures, but it is the original logic of Yule's derivation that I am after.

	Votes for Dukakis	Votes for Bush
Democratic Party	2	4
Republican Party	4	8

Data as Numbers

	Votes for Dukakis	Votes for Bush
Democratic Party	$n_{\text{Dem,Dukakis}}$	$n_{\text{Dem,Bush}}$
Republican Party	$n_{\text{Rep,Dukakis}}$	$n_{\text{Rep,Bush}}$

Data as Symbols

Figure 3.8
Data for Two Two-Valued Variables

Bush tells us nothing about the correlation between the two parties and the vote. For that we need to compare the Democrats to the Republicans. So Yule would have read the second message, again using odds, and then compared message to message by the ratio of their odds, which he named κ, the "odds ratio." In this example the odds ratio is 1, and now, with this ratio of two ratios, this comparison of comparisons, we know about the correlation between party and vote: In this example, there is no correlation. The logic of this correlation is summarized in Figure 3.9.

Yule used ratios, not differences, but his logic is clear enough.

Evidence:

Among Democrats:	Odds favoring Bush = 4/2
Among Republicans:	Odds favoring Bush = 8/4
Contrasting odds among Republicans to odds among Democrats:	Odds ratio, $\kappa = (8/4) / (4/2) = 1$

Conclusion:

"Vote," represented by the odds favoring Bush, is not correlated with Party: Comparing Republicans to Democrats, the odds do not change.

Figure 3.9
Logic of Contrasts for Two-Valued Variables

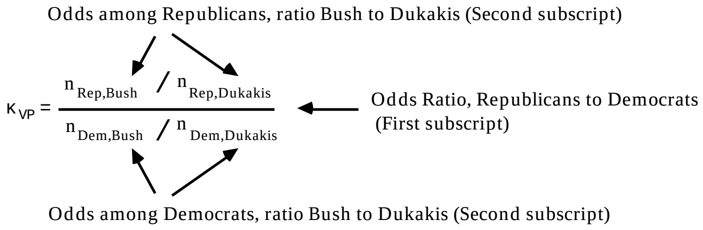

Odds among Republicans, ratio Bush to Dukakis (Second subscript)
$\kappa_{VP} = \frac{n_{Rep,Bush} / n_{Rep,Dukakis}}{n_{Dem,Bush} / n_{Dem,Dukakis}}$
Odds Ratio, Republicans to Democrats (First subscript)
Odds among Democrats, ratio Bush to Dukakis (Second subscript)

Figure 3.10
Yule's odds ratio, κ

Correlation lies in the double contrast measured by the ratio of odds. Visually, his logic is crystallized by the notation for κ, shown in Figure 3.10. In both the numerator and denominator the second subscript changes — expressing the comparison between Bush and Dukakis. Then, between the numerator and the denominator, the first subscript changes, expressing the comparison with respect to party.

The Conventions of Correlation: Yule's *Q*

Unfortunately, Yule too had a public that liked some kind of numbers and disliked others. The issue, in this case, was not a matter of subtraction, or division, or percentages. Rather, the issue was custom and custom in Yule's day, and our own, demanded that a measure of correlation satisfy three conventions. Logic aside, custom demanded:

1. That the value of a correlation coefficient would lie between +1 and –1 (or between +1 and 0 where "negative" correlation is undefined)

2. That the value of a correlation would be 0 for uncorrelated variables and

3. That the values of a correlation coefficient might change sign, but not magnitude, if the order of the rows or the columns were reversed.

Yule's odds ratio, κ, violated custom on all three counts: Violating the first convention, as a ratio, Yule's odds ratio lies between 0 and infinity, not plus-one and minus-one. Violating the second convention, Yule’s odds ratio is one for uncorrelated variables, not zero. And violating the third convention, the ratio changes from κ to 1/κ, not −κ, if the order of the rows or columns is reversed. Yule responded by transforming his κ to a new expression he named *Q*, bending and stretching κ to the demands of custom, using Equation 3.1. Yule's *Q*, named for Adolphe Quetelet, satisfied convention, but the logic was buried.

$$Q_{XY} = \frac{\kappa_{XY} - 1}{\kappa_{XY} + 1}, \qquad [3.1]$$

And, after ninety years of historical drift, the form of Yule's *Q* has itself been altered with the result that the presence of the odds ratios and the logic of Yule’s contrasts are entirely hidden, using the form in Equation 3.2. In theory, the original odds ratio and this modern expression of *Q* embody the same information. In practice there is a difference.

$$Q_{PV} = \frac{n_{RB} n_{DD} - n_{RD} n_{DB}}{n_{RB} n_{DD} + n_{RD} n_{DB}}. \qquad [3.2]$$

The Logic of Contrasts: Correlation and the Derivative

This transformation from κ to *Q* was, in a sense, a change of language — expressing the concept in the argot of statistics. But differences in language matter, the classical example being the difference between Roman numerals and Arabic positional numerals as two different expressions of numbers. Just try subtracting “C” from “XL” and then try subtracting “100” from “40.” One language works for science; the other is reserved for tombstones. In science language is active. It can block access to the logic of the science or simplify that access and accelerate it. And in this technical sense the conventional language of correlation, in Yule's day and in our own, is barbaric. Like Roman numerals, it labels the objects, but it inhibits their use by disguising the logic. It satisfies convention while obscuring the link between the principle of correlation and ideas of broad currency in the wider community of science.

Now, I'm ready to link correlation to the derivative. On the one side, correlation encodes a double contrast. On the other side, in the calculus there is a comparable double contrast: the rate of change with respect to x of the rate of change with respect to y (or, in the limit, the derivative with respect to x of the derivative with respect to y). If I link the two halves of the argument by using differences to express correlation — changing the language of correlation — then it becomes clear that these ideas are the same.

Using the difference, among Democrats, the contrast with respect to vote is the difference, Δ:

$$\begin{aligned}\Delta_{\text{Vote}}(\text{among Democrats}) &= \overset{\leftarrow \textit{Difference of log n with respect to Vote} \rightarrow}{[\,\log n_{\text{Dem,Bush}} - \log n_{\text{Dem,Dukakis}}\,]}\\ &= [\,\log 4 - \log 2\,]\,. \qquad [3.3]\end{aligned}$$

Using the difference, among Republicans the contrast with respect to votes is the difference:

$$\begin{aligned}\Delta_{\text{Vote}}(\text{among Republicans}) &= \overset{\leftarrow \textit{Difference of log n with respect to Vote} \rightarrow}{[\,\log n_{\text{Rep,Bush}} - \log n_{\text{Rep,Dukakis}}\,]}\\ &= [\,\log 8 - \log 4] \qquad [3.4]\end{aligned}$$

And then, using difference for the contrast with respect to party of these contrasts with respect to vote, Δ^2:

$$\Delta^{2}_{\text{Party,Vote}} = \overbrace{\overbrace{[\log n_{\text{R,B}} - \log n_{\text{R,D}}\,]}^{\substack{\textit{Difference of log n}\\ \leftarrow \textit{with respect to Vote} \rightarrow}} - \overbrace{[\log n_{\text{D,B}} - \log n_{\text{D,D}}\,]}^{\substack{\textit{Difference of log n}\\ \leftarrow \textit{with respect to Vote} \rightarrow}}}^{\substack{\textit{Difference with respect to Party}\\ \textit{of Differences of log n}\\ \leftarrow \textit{with respect to Vote-} \rightarrow}} \qquad [3.5]$$

There it is, "correlation" as the difference of two differences, the difference with respect to party of the differences with respect to vote. And to see why I find this exciting, let me juxtapose it with the derivative of the derivative — the mixed partial derivative. For well-behaved functions, the derivative of the derivative can be written in the form of Equation 3.6, which is itself the difference of two differences.

$$\frac{\partial^{2}}{\partial y\,\partial x}\, g(x,y) = \lim_{y \to y_o}\, \lim_{x \to x_o} \frac{\overbrace{\overbrace{[\; g(x,y) - g(x,y_o)\;]}^{\substack{\textit{Difference}\\ \leftarrow \textit{with respect to y} \rightarrow}} - \overbrace{[\; g(x_o,y) - g(x_o,y_o)\;]}^{\substack{\textit{Difference}\\ \leftarrow \textit{with respect to y} \rightarrow}}}^{\substack{\textit{Difference with respect to x}\\ \leftarrow \textit{of Differences with respect to y} \rightarrow}}}{(\,y - y_o)\,(\,x - x_o)} \qquad [3.6]$$

In Equation 3.5 you have correlation as a double contrast; in Equation 3.6 you have the mixed partial derivative as a double contrast and I submit that these two are the same. The terms within the square brackets of each expression are change with respect to one variable. And the difference between the terms in brackets is the change with respect to the other. Now I grant that describing these two equations as "the same" is sloppy: In the calculus, the expression "$x - x_o$" has numerical value where, in the table, there is only a difference between categories, like Bush and Dukakis. In fact, I'm willing to go all the way when I declare that these are the same — including the introduction of numerical values for all sorts of social science objects that have been presumed to be non-numerical, including Bush and Dukakis — but not in this chapter. For the moment I want to continue the single point that correlation is a disguised form of the basic concept of the derivative.

Gaussian, or "Normal," Correlation

So far I've attached my argument to a single measure of correlation, Yule's expression of correlation as *k* or *Q*. I'll admit that I'm using Yule: The test of the concept lies in its utility. And had Yule's example of correlation not illustrated my point, I would have chosen another example. But it was not necessary. Nor is it necessary in the case of another correlation, in the case of Gaussian, or "normal," correlation, ρ. If Yule's two-valued variables are the crudest variables that social science has to contend with, then Gaussian variables may be the most sophisticated, so sophisticated that they are not even real: In one dimension, Gaussian variables are the prime example of the bell-shaped curve, with values symmetrically distributed around their average value. In two-dimensions, Gaussian variables are the prime statistical model of a linear relation, with values of both variables symmetrically distributed around the line that describes their averaged relation. Gaussian variables are a statistical ideal, a mathematical norm to which the real world may or may not comply. For these idealized things there is an idealized correlation coefficient, ρ. And for ρ, as for *Q*, there is a direct link between correlation, Gaussian correlation, and the mixed partial derivative.

To show the correspondence, I am going to be intentionally abrupt: Never mind first principles of correlation that appear in any textbook on correlation. Never mind careful discussion of why the correspondence is reasonable. Instead, I am going to use the principle we have already developed, outlined in Figure 3.11, simply running the argument for ρ in parallel to the argument for *Q*: Thus, first transform the numbers. For the Gaussian this means using the logarithm of the Gaussian formula. Second, measure the change with respect to one of the two variables. For the Gaussian, this is the derivative. Third, measure the change with respect to the second variable of these changes with respect to the first. For the Gaussian, this is the mixed partial derivative. And that's the end of the argument because what's left shows ρ-correlation in a one-to-one correspondence with the mixed partial derivative. For ρ, as for *Q*, in each case the correlation coefficient carries the same information as the appropriate measure of change, stretched and shrunk to the limits of plus-one and minus-one, by the equations of Figure 3.12.[5]

[5] My apologies for word difficulties beyond my control: I am avoiding the conventional term "mixed partial derivative" and substituting "two-variable derivative." This is to avoid the word "partial" because partial has a different meaning in the context of correlation.

Page 1 of 2 page insert for Figure 3.11: Insert rotated 90 degrees relative to a normal page. These two should come out on facing pages.

Page 2 of 2 page insert for Figure 3.11: Insert rotated 90 degrees relative to a normal page.These two should come out on facing pages.

Q – Correlation Versus the Two-Variable Difference

ρ − Correlation Versus the Two-Variable Derivative

Correlation to Contrast

$$Q_{XY} = \frac{e^{\Delta^2_{XY}} - 1}{e^{\Delta^2_{XY}} + 1} \qquad \rho = \frac{-1 + (1+4\,g_{xy})^{\frac{1}{2}}}{2\,g_{xy}}$$

Contrast to Correlation

$$\Delta^2_{xy} = \frac{1 + Q_{XY}}{1 - Q_{XY}} \qquad Q_{XY} = \frac{e^{\Delta^2_{XY}} - 1}{e^{\Delta^2_{XY}} + 1}$$

where $\Delta^2_{xy} = \log \kappa$, the logodds ratio, and where $g(x,y) = \frac{\partial^2}{\partial y\ \partial x} g$

Graphically

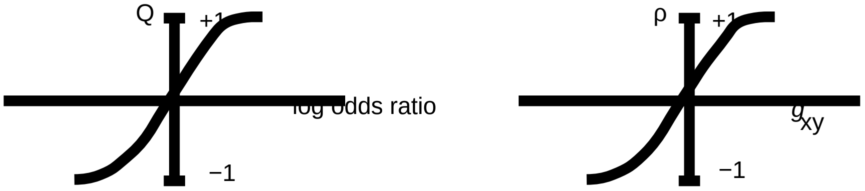


Figure 3.12
Transformations: Correlation to Change and Change to Correlation, Shown Algebraically and Graphically, for Both Q and ρ.

More-Variable Correlation

And there, in two instances, is Newton and Leibniz’s concept disguised, but functioning, as correlation. You discover the underlying rhythm, uniting our concept with theirs, by translation. And that's what I have to say about two-variable correlation, for the moment. Personally, I take pleasure in a concept that unifies ideas, ordering them and simplifying them. And the pleasure is all the greater when the unifying concept is already a classic, extending its reach. But we scientists are supposed to be practical too, and for that it is necessary to extend the concept beyond two variables to three variables and more.[6] Three-variable correlation is not in the vocabulary of statistics. There is three-variable linear regression, which is something else. And there is three-variable partial correlation, which is a special form of two-variable correlation, extracted from the complicating presence of other variables.

[6]. The subject is surely broader than I've indicated here and the parallels seem to continue for partial correlation. For partial correlation between Gaussian variables, the correct difference coefficient appears to be the average with respect to *z* of the derivative with respect to *x* and *y*. This average derivative corresponds to the usual "partial correlation" coefficient by exactly the same formula that translates the simple two-variable derivative into ρ.

For partial correlation between Yule's two-valued variables, you get something different in form from the partial *Q* suggested by Goodman and Kruskal but very similar in its numerical values when both are applied to real data. Goodman and Kruskal reasoned their way to a partial *Q* by working with the form of *Q* shown in Equation 3.2. Working, instead, with the difference form of Equation 3.5, and then translating from differences to kappa to *Q* leads to a partial *Q* that corresponds to the average odds ratio:

$$Q_{XY:Z} = \left(\bar{\kappa}_{XY:Z} - 1\right) \Big/ \left(\bar{\kappa}_{XY:Z} + 1\right)$$

where κ is the geometrical mean of the odds ratio, or the anti-log of the arithmetic mean of the difference of differences. In conventional form,

$$Q_{XY:Z} = \frac{n_{x'y'z'} n_{xyz'} n_{x'y'z} n_{xyz'} - n_{xy'z'} n_{x'yz'} n_{xy'z} n_{x'yz}}{n_{x'y'z'} n_{xyz'} n_{x'y'z} n_{xyz'} + n_{xy'z'} n_{x'yz'} n_{xy'z} n_{x'yz}} \; .$$

Both of these are interesting, but there are lots of things you can do with three variables. The narrower question is what do you get if you extend the logic of differences or derivatives to three variables and then use the result as correlation? The question is almost irresistible to anyone with the nasty turn of mind to push: to push for the next step, to push for more generality, to push to an extreme in which the argument will either fail or add insight. If this two-variable derivative corresponds to two-variable correlation, then what does the three-variable derivative or the three-variable difference correspond to? You have to ask what happens if you push the argument to more variables.

Precisely because we are using a calculus, the formulas for three-variable correlation are automatic. For the Gaussian it's very simple: There is none. Or, to be more precise, the three-way correlation is theoretically zero for Gaussian variables. By mathematical assumption the Gaussian has no three-variable "*xyz*" term that we would call three-way interaction, were it present. For real-world data such things may exist, but not for the Gaussian.[7]

For Yule's variables, however, Δ^3 is clear and not trivial. Logically, it must be the *difference* of the *differences* of the *differences*:

$$\Delta^3_{X\,YZ} = \Delta^2_{X\,Y}(z') - \Delta^2_{X\,Y}(z\,)\,. \qquad [3.7]$$

And now, simply invoking the equation that translates from Δ to κ and Q, Figure 3.12, you get the same concept expressed as a ratio, κ^3, and the same concept expressed as a Q, Q^3. What is it? As a ratio, κ^3 is the ratio of two odds ratios: It is the factor by which the odds ratio changes, as a function of a third variable, Z. As a Q it is a coefficient that juxtaposes opposite diagonals from the two different two-variable tables, multiplying the diagonal numbers, $n_{x'y'z'}n_{xyz'}$, from one table by the diagonal numbers, $n_{xy'z}n_{x'yz}$, from the other.

[7]. Specifically, no matter how many variables are included in the full *n*-variable Gaussian, it has no product terms with more than two variables.

$$Q_{XYZ} = \frac{\overbrace{n_{x'y'z'} n_{xyz'}}^{\text{Two Diagonal Elements from One } XY\text{-Table}} \overbrace{n_{xy'z} n_{x'yz}}^{\text{Two Diagonal Elements from the Other } XY\text{-Table}} - \overbrace{n_{xy'z'} n_{x'yz'}}^{\text{Two Off-Diagonal Elements from One } XY\text{-Table}} \overbrace{n_{x'y'z} n_{xyz}}^{\text{Two Off-Diagonal Elements from the Other } XY\text{-Table}}}{n_{x'y'z'} n_{xyz'} n_{xy'z} n_{x'yz} + n_{xy'z'} n_{x'yz'} n_{x'y'z} n_{xyz}} \qquad [3.\ 8]$$

This concept of three-variable correlation fits right in when you think in terms of differences — just add one more level of contrast. This "thing", implied by the logic, is what some social scientists call three-variable interaction, while others call it specification. Here it is shown as "simply" the three-variable generalization of correlation. It's certainly a known concept, but usually a negative one: When specification is present, ordinary two-variable methods fail, so we tend to watch for it and hope it isn't present. When specification is present we tend to separate the separate cases verbally, rather than quantitatively, and then discuss the separate two-variable cases.

I suspect that the limited use of the concept is caused, in part, by the lack of a number, the lack of a well-understood measure that would detect, express, and facilitate comparisons among three-way correlations — the way we use correlation to detect and compare different two-variable correlations.[8] Therefore, let me close this discussion by

[8] In both the analysis of variance and log-linear modeling there are tests for the presence of these effects. But just as we always caution users to separate the probabilistic "significance" of a two-variable correlation from the descriptive strength of the correlation, there is a difference between detecting the probabilistic significance of three-variable correlation and measuring its strength. Similarly, in both the analysis of variance and log-linear modeling we can measure the cell-by-cell consequences of correlation, but these effects are not the correlation

	Low GNP/Capita	High GNP/Capita
Low Literacy	51	15
High Literacy	14	50

$\Delta^2 = 2.50$, $\kappa = 12.14$, $Q = .84$

Figure 3.13
Relation Between Literacy and Per Capita Gross National Product

Low	High

coefficients. (The difference between the *effects* and the *correlation* is easily seen in a two-row, two-column table: The table has four interaction *effects* and one *correlation*.)

	Military Expenditures / GNP '78		Military Expenditures / GNP '78	
	Low GNP / Capita '75	High GNP / Capita '75	Low GNP / Capita '75	High GNP / Capita '75
Low Literacy '75	29	5	22	10
High Literacy '75	7	29	7	21

$\Delta^2 = 3.18$, $\kappa = 24.03$, $Q = .92$ $\Delta^2 = 1.89$, $\kappa = 6.60$, $Q = .74$

$\Delta^3 = 1.29$, $\kappa^3 = 3.6406926$, $Q^3 = .57$

Figure 3.14
Combined Data for Literacy and GNP Per Capita Separated According to High or Low Military Expenditures as a Fraction of GNP

applying the three-variable measure, derived from the model of the derivative, to data. For data, I've used various indicators for 155 nations, splitting each variable into two values, High and Low, accordingly as values were greater than the median for all values or less than or equal to the median.[9] Through all of these three-variable examples I've used two variables whose two-variable correlation is "obvious": Using literacy rates and per capita gross national products, the two-variable correlation is strong. Shown in Figure 3.13, nations with high literacy were twelve times more likely to show high GNP per capita. The odds were 50 to 14 for the more literate nations, compared to 15 to 51 for the less literate.

That's the way it's supposed to be — education leads to money and money leads to education, at least for a nation as a whole (nothing guarantees that the literate individuals in the country are the ones with the money). But, that's not the whole story. If military expenditures (military expenditures as a fraction of gross national product) are introduced as a third variable, then the *two*-variable correlations change: Where military expenditures are high, the correlation between literacy and wealth is relatively low. Where military expenditures are low, the correlation between literacy and wealth is relatively high.

Figure 3.14 shows the data. The correlations are positive in both sets of countries, but different. In fact, the presence or absence of high military expenditures effects a three- to four-fold change in the correlation between literacy and wealth. (The odds ratios change by a factor of 3.6.)

Habits with three variables are hard to break, so let's be clear what the data do *not* say about literacy, wealth, and military expenditures. The data do *not* say that high literacy and low military expenditures combine, or "add-up," to increase the wealth of the country. No, that would correspond to data in which the worst case for GNP would be low literacy and high military expenditures, while the best case would be the opposite, high literacy and low military expenditures. Lining up the

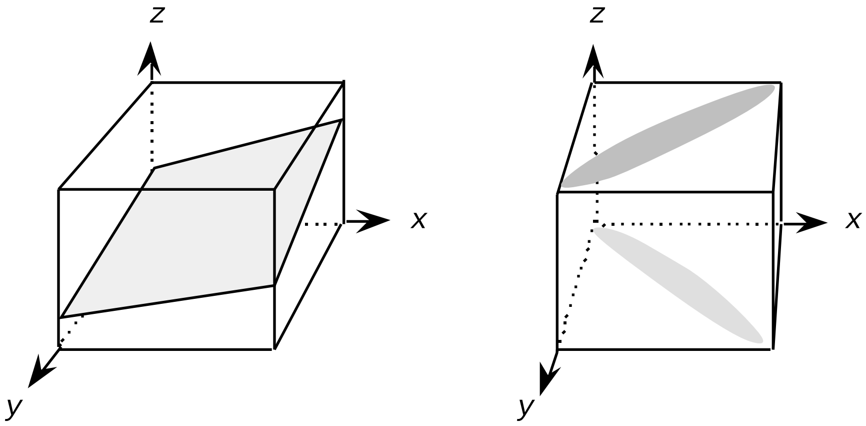
[9]. Using data reported in the *World Handbook of Political and Social Indicators*, Taylor and Jodice, 1983, available from the Interuniversity Consortium for Political and Social Research, Ann Arbor, Michigan.

Low Literacy High Literacy
←¾¾¾¾¾¾¾¾¾¾¾¾¾¾¾¾¾¾¾¾¾¾¾¾¾¾¾¾→
High Military Low Military

	Low Lit. High Mil.	Low Lit. Low Mil.	High Lit. High Mil.	High Lit. Low Mil.
Low GNP /C	22	29	7	7
High GNP /C	10	5	21	29
	Odds that GNP per capita will be High: 10 / 22 = .45	Odds that GNP per capita will be High: 5 / 29 = .17	Odds that GNP per capita will be High: 21 / 7 = 3.00	Odds that GNP per capita will be High: 29 / 7 = 4.14

Columns are ordered according to an additive model, with the order of the two middle categories being indeterminate. Counts indicate the number of countries displaying each combination. And the odds indicate the observed odds favoring High GNP per Capita

Figure 3.15
Comparing an "Additive Model" of the Odds Favoring High GNP per Capita (Left to Right) to the Actual Odds Favoring High GNP per Capita.



On the left the data lie in a plane; here the two-variable relations are constant, or uniform, throughout the data. On the right the data indicate a three-variable correlation; here, the two-variable relations change.

Figure 3.16
Two Forms of Three-Variable Relation

data in that order, in Figure 3.15, that is *not* what the data show. On the contrary, both the best case for high GNP and the worst case involve low military spending.

That's what is *not* in the data: In fact, the three-variable correlation is not a statement about wealth at all. It is a statement about the correlation between literacy and wealth: The correlation changes.

The difference between the usual use of three variables and the pattern detected by the three-variable correlation is shown graphically in Figure 3.16. On the left, the three-variable relation is, or is presumed to resemble, a plane: Within the plane the relation between *x* and *y*, between *x* and *z*, or between *y* and *z* is both linear and “constant,” the same throughout the data. On the right, the correlation changes. I’ve drawn it as a change in the *xy* correlation: It goes in one direction for low values of *z*; it goes in the opposite direction for high values of *z*. The two patterns, the one on the left and the one on the right, are quite incompatible: Real data can be matched by (at most) one of these patterns.[10]

The data for these nations can be set out in two more table formats, Figures 3.17 and 3.18, both of which show the same effect, although they are conducive to different interpretations of the data: As far as the numbers are concerned, it is equally correct to observe that literacy affects the correlation between wealth and military expenditures, Figure 3.17. Where literacy is high, there is a positive correlation between wealth and per capita military spending. Where literacy is low, there is a negative correlation. Comparing the less literate nations to the more literate nations the odds favoring military expenditures shift by a factor of 3.6, necessarily the same factor as for the first presentation of the data.

And finally, the data allow a third presentation, Figure 3.18, displaying the data in their full ambiguity. Separating the nations by GNP per capita, where wealth is low, there is a slight positive correlation between literacy and military spending. Where wealth is high, there is a negative correlation.

While all three of these arrangements are the same data, they look different, illustrating the positive effects of good form, the

[10]. The diagram is, of course, an exaggeration: The correlation must change, but it need not change sign. And while the figure on the right is drawn to show the change in the *xy* correlation, it is also true that three-variable correlation, or specification, changes the *xz* correlation and the *yz* correlation as well as the *xy* correlation.

One “obvious” candidate for a measure of three-variable correlation is a generalization of the two-variable correlation coefficient *r*. If you define r^3 as the mean product of three standardized variables (standardized to mean 0 and standard deviation 1), then the result seems responsive to the presence of these three-variable correlations. But this r^3 is not satisfactory without modification because it is also responsive to data distributions that are not indicative of three-variable correlation: It achieves its maximum values for variables that are co-linear, and highly skewed, which is not what I’m after. (It is not bounded by plus- and minus-one, but that’s a minor problem.)

	Low Literacy '75		High Literacy '75	
	Low Military Expenditures /GNP '78	High Military Expenditures/GNP '78	Low Military Expenditures/GNP '78	High Military Expenditures/GNP '78
Low GNP/ Capita '75	29	22	7	7
High GNP/ Capita '75	5	10	29	21

$\Delta^2 = .97$, $\kappa = 2.64$, $Q = .45$ $\Delta^2 = -.32$, $\kappa = .72$, $Q = -.16$

$\Delta^3 = 1.29$, $\kappa^3 = 3.64$, $Q^3 = .57$

Figure 3.17
Combined Data for GNP Per Capita and Military Expenditures as a Fraction of GNP, Separated according to High or Low Literacy

	Low GNP/Capita '75		High GNP/Capita '75	
	Low Military Expenditures /GNP '78	High Military Expenditures/GNP '78	Low Military Expenditures/GNP '78	High Military Expenditures/GNP '78
Low Literacy '75	29	22	5	10
High Literacy '75	7	7	29	21

$\Delta^2 = .28$, $\kappa = 1.32$, $Q = .14$ $\Delta^2 = -1.02$, $\kappa = .36$, $Q = -.47$

$\Delta^3 = 1.29$, $\kappa^3 = 3.64$, $Q^3 = .57$

Figure 3.18
Combined Data for Literacy and Military Expenditures as a Fraction of GNP, Separated according to High or Low GNP per Capita

negative effects when it's absent: Using Δ's to measure correlation it is clear that the differences of correlation are the same, 1.29 in each case.

Even using κ's, it is almost clear that the change of odds ratios, from 24 to 6.6, from 2.6 to .7, and from 1.3 to .4, is the same in each case. Using *Q*s, satisfying convention, the information is present but obscure, .92 versus .74, .45 versus –.16, and .14 versus –.47. "Everyone" knows it's hard to compare two *Q*s, but it's also hard to resist. In the first presentation, the two *Q*s look "large." They look approximately equal. In the second and third presentations the two *Q*s differ in sign and magnitude. They look different. It would be easy to miss the specification in one case or focus on it in the other, using *Q*. Reverting to good form, using differences, these are all the same.

Scanning the social and political indicators for other instances of three-variable correlation yields a group of "third" variables suggesting, jointly, that the correlation between literacy and wealth (between literacy rates and per capita gross national products) is specified by the size of the country: In small countries, the correlation is relatively large, or more positive. In large countries the correlation between literacy and wealth is relative small, or less positive. The specifying variables include three measures of population, one measure of physical area, and one more measure of expenditure.

Specified by

Total Adult Population:	$\Delta^3 = 0.86$,	$\kappa^3 = 2.36$,	$Q^3 = .40$
Total Military Manpower:	$\Delta^3 = 0.90$,	$\kappa^3 = 2.47$,	$Q^3 = .42$
Total Population:	$\Delta^3 = 0.98$,	$\kappa^3 = 2.67$,	$Q^3 = .46$
Total Agricultural Area:	$\Delta^3 = 1.10$,	$\kappa^3 = 3.00$,	$Q^3 = .50$
Total Working Age Population:	$\Delta^3 = 1.22$,	$\kappa^3 = 3.40$,	$Q^3 = .55$
Total Defense Expenditures:	$\Delta^3 = 1.24$,	$\kappa^3 = 3.46$,	$Q^3 = .55$

Pushing Newton and Leibniz's concept to three variables, these six correlations indicate differences in the correlation between literacy and wealth. These correlation coefficients are Newton and Leibniz's concept at work, disguised as correlation.

Appendix 3.1

1988 Population in thousands and 1990 Gross National Product in millions of dollars. For example, the United States data indicate 251 million people and 4.86 trillion dollars. Data from 1991 *Britannica Book of the Year*, Encyclopaedia Britannica, Inc., Chicago, 1991.

	Population	GNP		Population	GNP
	in	in		in	in
	Thousands	Millions		Thousands	Millions
		of Dollars			of Dollars
United States	251,394	4,863,674	Argentina	32,880	83,040
Japan	123,530	2,576,541	Romania	23,265	79,813
Soviet Union	290,122	2,500,000	South Africa	37,418	77,720
Germany (West)	62,649	1,131,265	Indonesia	180,763	75,960
France	56,647	898,671	Turkey	56,941	68,600
Italy	57,512	765,282	Hungary	10,437	64,527
United Kingdom	57,384	730,038	Venezuela	19,735	59,390
Canada	26,620	437,471	Algeria	25,337	58,250
China	1,133,683	356,490	Hong Kong	5,841	54,567
Brazil	150,368	328,860	Thailand	56,217	54,550
Spain	38,959	301,829	Bulgaria	8,997	50,837
India	853,373	271,440	Greece	10,038	48,040
Netherlands	14,934	214,458	Iraq	17,754	40,700
Australia	17,073	204,446	Israel	4,666	38,440
Switzerland	6,756	178,442	Philippines	61,480	37,710
Poland	38,217	172,774	Portugal	10,388	37,260
Sweden	8,529	160,029	Colombia	32,978	37,210
Germany (East)	16,433	159,370	Pakistan	122,666	37,153
Mexico	81,883	151,870	Egypt	53,170	33,250
Korea (South)	42,793	150,270	New Zealand	3,389	32,109
Belgium	9,958	143,560	Nigeria	119,812	31,770
Yugoslavia	23,800	129,514	Malaysia	17,886	31,620
Taiwan	20,221	125,408	Peru	22,332	29,185
Czechoslovakia	15,664	123,113	Cuba	10,603	26,920
Austria	7,623	117,644	Ireland	3,509	26,750
Denmark	5,139	94,792	Kuwait	2,143	26,250
Iran	56,293	93,500	Singapore	2,718	24,010
Finland	4,978	92,015	United Arab Emir	1,903	23,580
Saudi Arabia	14,131	86,527	Libya	4,206	23,000
Norway	4,246	84,165	Korea (North)	22,937	20,000

	Population	GNP		Population	GNP
	in	in		in	in
	Thousands	Millions		Thousands	Millions
		of Dollars			of Dollars
Syria	12,116	19,540	Bolivia	7,322	3,930
Chile	13,173	19,220	Tanzania	24,403	3,780
Puerto Rico	3,336	18,520	Mongolia	2,116	3,620
Bangladesh	113,005	18,310	Gabon	1,171	3,200
Morocco	25,113	17,830	Afghanistan	15,592	3,100
Vietnam	66,128	12,600	Brunei	259	3,100
Cameroon	11,900	11,270	Bahrain	503	3,027
Ecuador	10,782	10,920	Papua New Guinea	3,671	2,920
Tunisia	8,182	9,610	Nicaragua	3,871	2,911
Cote d'Ivoire	12,657	8,590	Nepal	18,910	2,843
Luxembourg	378	8,372	Bahamas	253	2,611
Kenya	24,872	8,310	Macau	461	2,611
Sudan	28,311	8,070	Jamaica	2,391	2,610
Guatemala	9,197	7,620	Guinea	6,876	2,300
Myanmar	41,675	7,450	Haiti	5,862	2,240
Uruguay	3,033	7,430	Niger	7,779	2,190
Oman	1,468	7,110	Zambia	8,456	2,160
Sri Lanka	17,103	7,020	Madagascar	11,980	2,080
Angola	10,002	6,930	Rwanda	7,232	2,064
Zimbabwe	9,369	6,070	Burkina Faso	9,012	1,960
Ethiopia	50,341	5,760	Congo	2,326	1,950
Zaire	34,138	5,740	Mauritius	1,080	1,890
Yemen (San'a')	9,060	5,700	Reunion	600	1,830
Ghana	15,020	5,610	Lebanon	2,965	1,800
Panama	2,418	5,091	Mali	8,152	1,800
Iceland	256	5,019	Malta	353	1,740
El Salvador	5,221	4,780	Jersey	83	1,647
Paraguay	4,279	4,780	Mozambique	15,696	1,550
Costa Rica	3,015	4,690	Barbados	257	1,530
Dominican Rep	7,170	4,690	Benin	4,741	1,530
Senegal	7,317	4,520	West Bank	908	1,500
Uganda	16,928	4,480	Namibia	1,302	1,477
Jordan	3,169	4,420	Bermuda	59	1,406
Cyprus	739	4,320	Martinique	261	1,400
Trinidad&Tobago	1,233	4,160	Fr. Polynesia	197	1,370
Honduras	4,674	4,110	Malawi	8,831	1,320
Qatar	444	4,060	Togo	3,764	1,240
Albania	3,262	4,030	Burundi	5,451	1,200

	Population in Thousands	GNP in Millions of Dollars
Botswana	1,295	1,191
Guadeloupe	380	1,170
Fiji	740	1,130
Guernsey	60	1,122
Cen. African Rep	2,875	1,080
Virgin Islands	105	1,070
Liberia	2,595	1,051
Suriname	411	1,050
Guam	132	1,000
Yemen (Aden)	2,486	1,000
Somalia	7,555	970
Sierra Leone	4,151	930
Mauritania	1,999	910
Neth. Antilles	196	860
New Caledonia	168	856
Chad	5,678	850
Laos	4,024	710
Lesotho	1,760	690
Faeroe Islands	48	686
Aruba	63	619
Cambodia	8,592	600
Swaziland	770	580
Gaza	608	560
Greenland	56	465
Cayman Islands	26	461
Liechtenstein	29	450
Andorra	51	360
Isle of Man	64	340
Djibouti	530	330
Guyana	756	327
Monaco	29	280
Belize	189	264
Seychelles	69	260
Nor. Mariana Is	23	256
Antigua	81	230
St. Lucia	151	220
Bhutan	1,442	202
Comoros	463	200
American Samoa	40	190
San Marino	23	188
Gambia	860	180
French Guiana	117	176
Cape Verde	339	170
Nauru	9	160
Guinea-Bissau	973	145
Equator. Guinea	350	140
Grenada	101	139
Br. Virgin Is.	13	133
Dominica	82	130
Gilbraltar	31	130
Solomon Islands	319	130
St. Vincent	115	130
St. Kitts-Nevis	44	120
Vanuatu	147	120
Micronesia	108	107
Western Samoa	165	100
Maldives	214	80
Tonga	96	80
Turks & Caicos	15	63
Falkland Islands	2	56
Montserrat	12	54
Marshall Islands	46	46
Kiribati	71	40
Palau	14	32
SaoTome&Principe	121	32
Anguilla	7	28
Cook Islands	19	21
Wallis & Futuna	16	10
Niue	2	3
Tuvalu	9	3
Tokelau	2	1

Appendix 3.2

Three-variable data, indicating the names as well as the counts of the nations showing each combination. These are the combined data for Literacy, for GNP Per Capita, and for Military Expenditures as a Fraction of GNP, comparable to Figure 3.14.

	Low Military Expenditures/GNP '78		High Military Expenditures/GNP '78	
	Low GNP/Capita '75	High GNP/Capita '75	Low GNP/Capita '75	High GNP/Capita '75
Low Literacy '75	29 Haiti, Guatemala, Honduras, Gambia, Senegal, Benin, Niger, Ivory Coast, Guinea, Liberia, Sierra Leone, Ghana, Togo, Cameroon, Central African Republic, Zaire, Kenya, Burundi, Rwanda, Ethiopia, Mozambique, Malawi, Lesotho, Swaziland, Madagascar, Afghanistan, Bangladesh, Nepal, Papua New Guinea	5 Nicaragua, Gabon, Algeria, Tunisia, Libya	22 Cape Verde, Guinea Bissau, Mali, Mauritania, Upper Volta, Nigeria, Chad, Congo, Uganda, Somalia, Zambia, Zimbabwe, Botswana, Morocco, Sudan, Egypt, Jordan, Yemen Sana, Yemen Aden, India, Pakistan, Laos	10 South Africa, Iran, Turkey, Iraq, Syria, Saudi Arabia, Kuwait, Bahrain, United Arab Emirates, Malaysia
High Literacy '75	7 El Salvador, Columbia, Ecuador, Bolivia, Paraguay, Sri Lanka, Philippines	29 Canada, Dominican Republic, Jamaica, Trinidad Tobago, Barbados, Mexico, Costa Rica, Panama, Venezuela, Surinam, Brazil, Chile, Argentina, Uruguay, Ireland, Luxembourg, Switzerland, Spain, Austria, Italy, Malta, Cyprus, Finland, Denmark, Iceland, Mauritius, Japan, Australia, New Zealand	7 Guyana, Tanzania, China, Korea South, Burma, Thailand, Indonesia	21 United States, Cuba, Peru, United Kingdom Netherlands, Belgium, France, Portugal, Federal Republic Germany, German Democratic Republic, Poland, Hungary, Yugoslavia, Greece, Bulgaria, Romania, USSR, Sweden, Norway, Israel, Singapore

$$\Delta^3 = 1.29,\ \kappa^3 = 3.6406926,\ Q^3 = .57$$