

Best Practices for Understanding Quantitative Data

Jonathan G. Koomey, Ph.D.

February 14, 2006

Anyone who has delved into data from the real world knows it can be messy. Survey takers can write down the wrong response. People entering data into a computer can type in the wrong numbers. Computers sometimes garble data because of software bugs (especially when converting one file format to another). Data formats become obsolete as software changes. Electronic data recorders break or go out of adjustment. Analysts mislabel units and make calculational errors.

Cleaning the data to correct such problems is almost always necessary, but this step is often ignored. Ask about the process that was used to clean the data. If the response is just a blank stare or worse, you know you are in trouble. Many companies (such as AT&T and Sega of America) now assign people to check for data quality in the face of all the problems associated with real data.

Some of the more famous examples of how small mistakes in data processing can have disastrous results are found in the space program. In 1962, the Mariner I spacecraft went off course soon after launch and had to be destroyed. The cause of the malfunction was a missing hyphen in a single line of FORTRAN code. Arthur C. Clarke later quipped that this \$80 million launch was ruined by “the most expensive hyphen in history.” ([CNET](#) online retells the story of this launch as well as other famous computer glitches.)

More recently (1999), NASA’s Mars Climate Orbiter was lost in space because engineers on the project forgot to convert English units to metric units in a key data file, a mistake that cost scientists years of work and taxpayers \$125 million. (Source: Sawyer, Kathy. “Engineers’ Lapse Led to Loss of Mars Spacecraft: Lockheed Didn’t Tally Metric Units.” *The Washington Post*. October 1, 1999, p. A1.)

Bad data can have a real cost for a business. If a 500,000-piece mailing uses an address list with an error rate of 20%, the company wastes \$300,000 by sending mailings to incorrect addresses. It loses even more money because among those 100,000 missed prospects are about 1,500 people who would have become customers and purchased thousands of dollars worth of the company’s products over their lifetime. The losses from these missed customers can be many times greater than the immediate direct losses. (Source: Aragon, Lawrence. “Down with Dirt.” *PC Week*. October 27, 1997, p. 83.)

It is crucial to pore over raw data to check for anomalies before doing extensive analyses. For example, typographical errors can lead numbers to be ten, one hundred or one million times bigger than they should be. Looking over the raw data can help you identify such problems before you waste time doing analysis using incorrect numbers.

Bad data can ruin your credibility and call your work into question. Even if there is only one small mistake, it makes your readers wonder how many other mistakes have crept into your analysis. It is difficult to restore your credibility after some obvious mistake is revealed, so avoid this problem in the first place. Dig into your numbers and root out these problems before you finalize your memo, paper or presentation.

Some Specific Advice

If your data come from a non-electronic source, type the data into the computer yourself, assuming there is a manageable amount of data. There is no substitute for this effort, even if you are a highly paid executive, because it will help you identify inconsistencies and problems with the data and give you ideas for how to interpret them. This technique also gives you a feeling for the data that cannot be replicated in any other way. You will almost surely see patterns and gain unexpected insights from this effort.

Check that the main totals are the sum of the subtotals. Most documents are rife with typographical errors and incorrect calculations. Therefore, you should not rely blindly on any data source's summations but calculate them from the base data. You can check your typing accuracy by comparing the sums to those in the source of data. If they match exactly, it is unlikely that your typing is in error. Even if you don't check these sums, you can bet that some of your readers or listeners will. Do it yourself and avoid that potential embarrassment.

Check that the information is current. Do not forget that business and government statistics are revised regularly. Make sure you know the vintage of the input data used in the analysis. For example, don't compare analysis results generated using one year's census data with those based on another year's data (unless your sole purpose is to analyze trends over time).

Check relationships between numbers that should be related in a predictable way. Such comparisons can teach valuable lessons. For example, when examining data on carbon emissions of different countries, a newcomer to the field of greenhouse gas emissions analysis might expect that the amount of carbon emitted per person would not differ much among industrialized countries. In examining such data, however, we find large differences in carbon emitted per person, from less than 1 metric ton/person/year in Portugal to more than 6 tons/person/year in Luxembourg. Determining why such differences exist is the logical next step, which will inevitably lead to further analysis and understanding.

Check that you can trace someone else's calculation in a logical way. If you cannot do this, you can at least begin listing the questions you need to answer to start tracing the calculation. Ultimately, if you cannot reproduce the calculation, the author has broken a fundamental rule of good data presentation, and his analysis is suspect.

Compare the numbers to something else with which you are familiar, as a "first-order" sanity check. These comparisons can show you whether or not you are on the right track. Presenting such comparisons in reports and talks can also increase your credibility with your readers or listeners because it shows that your results "pass the laugh test."

Normalize numbers to make comparisons easier. For example, the true size of total U.S. gross national product (GNP) in trillion dollars per year is difficult to grasp for most people, but if normalized to dollars per person per year will be a bit more understandable. Common bases for such normalizations are population (per person/per capita), economic activity (per dollar of GNP), or physical units of production (per kilowatt hour or per kilogram of steel produced).

If you have information that extends over time ("time series data"), normalize it to a base year to enhance comparisons. By expressing such data as an index (e.g., 1940 = 1.0), you can

compare trends to those of other data that might be related. For example, if you plot U.S. raw steel production (Figure 1), population (Figure 2), and GNP (Figure 3) in separate graphs, it is difficult to gain perspective on how fast steel production is changing over time relative to these other two important determinants of economic and social activity. However, if you plot steel production over time as an index with 1940 = 1.0 (see Figure 4), you can plot population and GNP on the same graph. Such a graph will instantly show whether growth rates in the data differ.

In this example, real GNP grew by a factor of more than five from 1940 to 1990. U.S. steel production roughly doubled by 1970 and then declined by 1990 to roughly 50% above 1940 levels. The population in 1990 just about doubled from 1940 levels. (Sources: 1940-1980 GNP from the *Statistical Abstract of the US 1990*, p. 425. 1990 GNP in current dollars from 1997 *World Almanac and Book of Facts*, p. 133, adjusted to 1982 dollars using the consumer price index from p. 132 of that document.)

The trends for U.S. steel production in Figure 4 dramatically illustrate the changing fortunes of steel in a postindustrial economy. Just after World War II, steel use per capita increased as automobile ownership expanded and use of steel for bridges and other forms of construction also increased. After 1970, the steel industry began to face serious competition from foreign steel makers as well as from alternative materials such as aluminum alloys and composites. Consequently, U.S. production declined even as the population increased and real GNP went through the roof. (Source: U.S. raw steel production from 1997 *World Almanac and Book of Facts*, p. 153.)

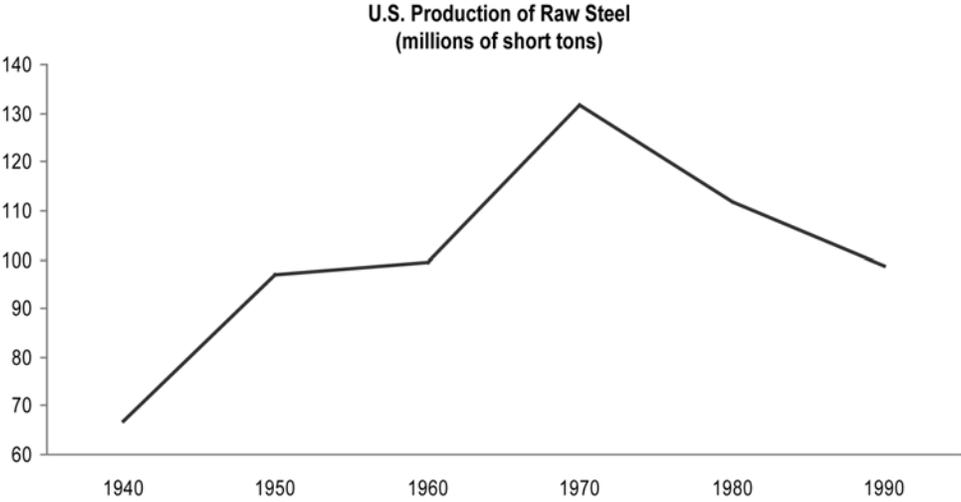


Figure 1: U.S. production of raw steel 1940-1990 (million short tons)



Figure 2: U.S. population 1940-1990 (million people)

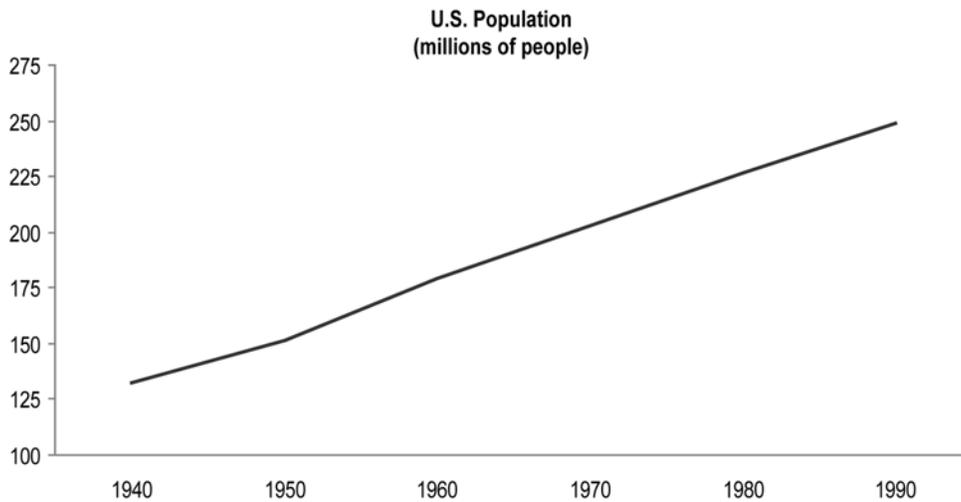


Figure 3: U.S. gross national product 1940-1990 (billion 1982 dollars)

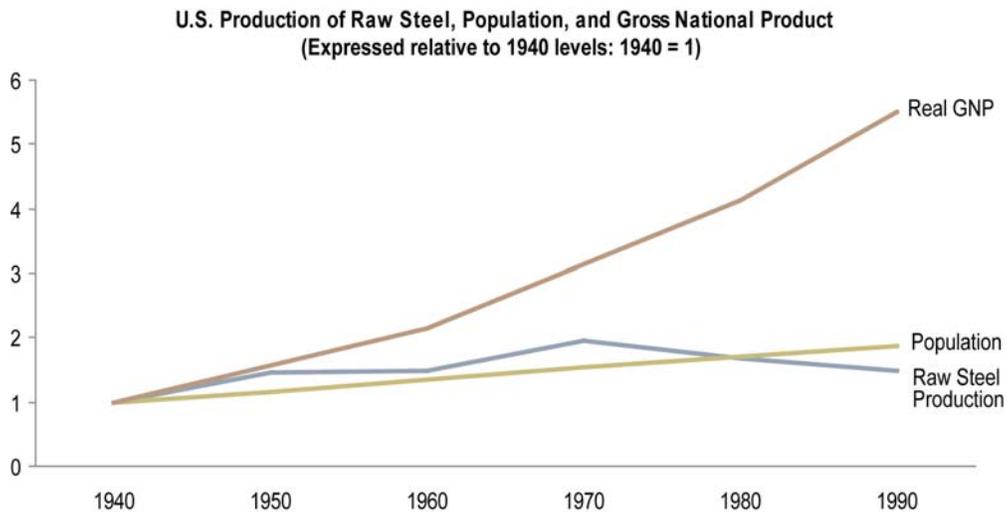


Figure 4: U.S. raw steel production, population, and gross national product 1940-1990, expressed relative to 1940 levels (1940 = 1)

Break problems into component parts. Explore analysis results by examining the factors that led to those results. For example, suppose someone tells you that the market capitalization of Google in January 2006 was about \$130 billion and that of General Electric (GE) was \$350 billion. What steps should you take to understand what these numbers mean?

Market capitalization is the product of the number of shares outstanding and the stock price per share, as shown in the following equation:

$$\text{Market Capitalization (\$)} = \# \text{ Shares} \times \frac{\text{Stock Price}}{\text{Share}}$$

The stock price per share can be further broken down into the product of the earnings per share and the price-to-earnings ratio, yielding the following simple model:

$$\text{Market Capitalization (\$)} = \# \text{ Shares} \times \frac{\text{Earnings}}{\text{Share}} \times \frac{\text{Price}}{\text{Earnings}}$$

The product of the number of shares and the earnings per share gives the total annual earnings (profits) for each company. Substituting in the previous equation, we get:

$$\text{Market Capitalization (\$)} = \text{Total Earnings} \times \frac{\text{Price}}{\text{Earnings}}$$

If we divide both sides of this equation by annual revenues, we get:

$$\frac{\text{Market Capitalization}}{\text{Annual Revenues}} = \frac{\text{Total Earnings}}{\text{Annual Revenues}} \times \frac{\text{Price}}{\text{Earnings}}$$

All of these equations represent variations on the same model. Different forms of the model will be useful at different times.

For most companies, the basic information for the model is readily available on the Web, so that is the best place to start (you could also go to the library). Table 1 summarizes the key financial parameters for calculating market capitalization for the two companies (taken from Yahoo Finance on January 29, 2006).

The first thing to notice about the financial statistics for these two companies is a huge disparity. While GE's market capitalization in January, 2006, was about three times larger than Google's, GE's revenues were almost thirty times larger. All other things being equal, we might expect that companies with similar market valuations would also have similar revenue streams. All other things are not equal, however, and finding out why will help illustrate this important analytical technique.

If Google's market capitalization per dollar of revenues were the same as GE, we would expect that its market capitalization would be only one-tenth as large as it is. We need to explain this tenfold discrepancy. To do so, we examine the components of the last equation above. The first component is earnings per dollar of revenues, and the second component is the price-to-earnings ratio.

As Table 1 shows, Google's earnings in January, 2006, were about twice as big as GE's per dollar of revenues, which accounts for about a factor of 2.3 in our tenfold difference. The price-to-earnings ratio also differed between the two companies. Apparently, the stock market valued one dollar of Google's earnings 4.5 times as much as one dollar of GE's earnings, which accounts for the remaining difference.

Table 1: Comparison of financial statistics for General Electric and Google

| | <i>Units</i> | <i>GE</i> | <i>Google</i> | <i>Ratio Google /GE</i> |
|---------------------------------------|--------------|-----------|---------------|-----------------------------|
| Annual revenues | B\$ 2006 | 148 | 5.3 | 0.035 |
| Number of shares | Billions | 11 | 0.30 | 28 |
| Price per share | \$2006/share | 33 | 433 | 13.2 |
| Diluted Earnings per share (EPS) | \$2006/share | 1.54 | 4.51 | 2.9 |
| Trailing price-to-earnings (PE) ratio | Ratio | 21 | 96 | 4.5 |
| Earnings per dollar of revenues | \$/dollar | 0.11 | 0.25 | 2.3 |
| Market capitalization | B\$ 2006 | 348 | 128 | 0.37 |
| Market cap. per dollar of revenues | \$/dollar | 2.4 | 24 | 10.4 |

(1) B signifies billions (thousand millions for British readers).

(2) PE is the share price to earnings ratio (measured as an intraday value).

(3) Market capitalization (measured as an intraday value) is the product of # of shares and price per share. Price per share is the product of EPS and PE ratio.

(4) Values taken from Yahoo Finance on January 29, 2006.

By breaking these numbers into their component parts, we have been able to isolate the two key reasons for the tenfold discrepancy identified above. The next step is to explain why these two reasons exist.

Google's greater profitability per dollar of revenue reflects the difference between industrial manufacturing and internet software development. The latter has very low marginal costs of reproducing the product and high gross margins. Google's higher price-to-earnings ratio reflects the market's belief that its dominance will allow the company to continue to generate growth in earnings at a pace vastly exceeding that of traditional industrial enterprises.

Dissecting analysis results by comparing ratios of key parameters is a powerful approach, and one to use frequently. Any time you have two numbers to compare, this kind of "ratio analysis" can lead to important insights into the causes of underlying differences between the two numbers.

Creating a presentation or an executive summary for a complex report always involves boiling an analysis down to the essentials. First, create equations (like the ones above) that calculate key analysis results as the product of several inputs multiplied together. Then, determine which of these inputs affects the results most significantly. Creating such models can help you think systematically about your results.

Applying these Skills to Reading Tables and Graphs

The profusion of tables and graphs in magazine and news articles gives you many opportunities to practice these skills, which will help you understand the bottom-line results and determine whether you find the author credible. If the tables and graphs are good enough, you can then read the paper to follow the author's reasoning more closely.

If the tables and graphs are poorly designed or confusing, I lose respect for the author. It is essential that tables and graphs summarizing analysis results be clear, accurate, and well documented. If they aren't, including them is worse than useless because they hurt your argument and your credibility.

Start by checking for internal consistency. I always begin at the bottom line of the table and work backward. I examine the column and row headings to be sure I understand what each one represents (I read the footnotes if I have questions). Then, I assess whether the components of the total add up to the total. This procedure shows me whether the calculations are accurate, and it helps me become familiar with the various parts of the analysis.

Not surprisingly (but fortuitously for purposes of this article), I found one internal inconsistency in the data on Yahoo finance in the course of creating the comparison between Google and GE above. Yahoo gives revenues per share of \$19.6 for Google, but if you multiply those revenues per share by the number of shares, you get total revenue of \$5.8 billion, instead of the \$5.25 billion revenues listed in Yahoo finance. I assumed that the total revenues and numbers of shares given by Yahoo were correct and adjusted the revenues per share to reflect that assumption. You can't take any data for granted!

It is a good idea to look over the numbers in the table and identify those that are abnormally small or large. Typographical errors are quite common in tables (particularly in tables that summarize results from other, more detailed tables), and a quick scan can help you find them. For example, if you are reading a table summarizing hours worked per week by different team members on a project, an entry from one person that is ten times larger than the entries for others should catch your attention and prompt you to investigate further. The number itself may not be wrong, but checking it will increase your confidence in the numbers and help you understand how they were calculated.

Sometimes numbers do not exactly add up because of *rounding errors*, not because there is a mistake in the calculations. For example, say you have formatted a spreadsheet table so there are no decimal places for the entries in the table. These entries might be 9.4 and 90.4, but in the table they are shown as 9 and 90 because the convention is to round numbers down to the next whole number when the decimal remainder is less than 0.5 (the remainder in both cases is 0.4 in this example). The sum is 99.8, which rounds to 100 and is shown in the spreadsheet as the total. The sum of 9 and 90 is 99, which makes the total of 100 look wrong even though there is a perfectly sensible explanation. If you're not aware of this potential pitfall, you could be misled by these apparent errors.

Read the footnotes carefully. They should convey the logic of the calculations and the sources of key input data. If you cannot determine the methods used from the footnotes, you should be especially suspicious of the results and investigate further.

Check for ambiguous definitions and terminology. For example, there are at least five distinct definitions of the word “ton,” and analysts often neglect to specify which definition they are using. If it is not crystal clear what the label means, you are likely to be led astray when interpreting the numbers. A slightly different example involves the number of hours in a year. Many analysts assume it is 8,760 hours, but on average it is 8,766 hours because of leap years. This difference is essentially a definitional one, but it can lead to small inaccuracies in calculations.

The next step is to check consistency with independent external sources to make sure the values in the tables and graphs are roughly right. As described earlier, compare growth over time to growth in other key data, such as population and gross domestic product to get perspective on how fast something is growing relative to these commonly used indicators.

Does the information in the tables or graphs contradict other information you know to be true? That table of hours worked may list Joe as someone who worked little on the project at hand; but if you know for a fact that Joe slaved over this project for many weeks because it was his idea, then you will need to check the calculation. Similarly, if there is an entry of 175 hours for one week, it must be a typo or a miscalculation because there are only 168 hours in a week.

Take ratios of results and determine whether the relationships they embody make sense. If one component of the total is growing much faster than other components over time or if it is especially large compared to others, then investigate further. Look for large discrepancies and investigate when you uncover them.

Follow up when you encounter cognitive dissonance—any contradiction between your knowledge and the information in the table will lead to greater understanding, one way or the other. If there is a logical explanation for the contradiction, you have learned more about the relationships between the information in the table and what you knew before. If the contradiction indicates a real inconsistency, you have identified a flaw in the analysis. Root out the causes of cognitive dissonance, and you will enhance your knowledge without fail.

Conclusions

When John Holdren was a professor at the University of California Berkeley, he taught a delightful class titled “Tricks of the Trade.” In this class, he described many of the unwritten rules about being effective in the energy/environment field and listed key pitfalls in data acquisition and handling. I have aggregated them below into four golden rules:

- *Avoid information that is mislabeled, ambiguous, badly documented, or otherwise of unclear pedigree.* Ambiguity and poor documentation are an indication that the quality control for such data is uneven at best and appalling at worst. Dig into the numbers a bit and find out whether it is carelessness or incompetence; make sure you believe the numbers before using them.
- *Discard unreliable information that is invented, cooked, or incompetently created.* If you find major inconsistencies, conceptual flaws, and omissions in the data, discard that information, no matter how much it might help your analysis.
- *Beware of illusory precision.* Do not represent or interpret information as more accurate than it is. Carefully characterize uncertainty and variability in your data, and insist that others do so with their own.

- *Avoid spurious comparability.* Beware of numbers that are ostensibly comparable but fundamentally inconsistent. Create and use only consistent comparisons.

Holdren's advice when dealing with data is: "Be suspicious, skeptical, and cynical. Assume nothing." Though it may sound paranoid to the uninitiated, such caution is an absolute necessity for the seasoned business analyst.

About the Author

Jonathan G. Koomey, Ph.D. is a Staff Scientist at Lawrence Berkeley Laboratory and a Consulting Professor at Stanford University. He is the author or co-author of eight books and more than one hundred and fifty articles and reports on the economics of energy technologies, environmental policy and critical thinking skills. Jon holds M.S. and Ph.D. degrees from the Energy and Resources Group at the University of California at Berkeley, as well as an A.B. in History of Science from Harvard University. This article is adapted from Chapters 18 and 27 of Koomey's book *Turning Numbers into Knowledge: Mastering the Art of Problem Solving*, Analytics Press: Oakland, CA 2004.

This article originally appeared on the Business Intelligence Network (www.b-eye-network.com) as one of Stephen Few's guest articles. A library of Stephen Few's articles, as well as other guest articles, is available at www.perceptualedge.com.