



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 in-

formation 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.