Data Analysis

By Robert Niles

You wouldn't buy a car or a house without asking some questions about it first. So don't go buying into someone else's data without asking questions, either.

Okay, you're saying... but with data there are no tires to kick, no doors to slam, no basement walls to check for water damage. Just numbers, graphs and other scary statistical things that are causing you to have bad flashbacks to your last income tax return. What the heck can you ask about data?

Plenty. Here are a few standard questions you should ask any human beings who slap a pile of data in front of you and ask you write about it.

1. **Where did the data come from?** Always ask this one first. You always want to know who did the research that created the data you're going to write about.

   You'd be surprised - sometimes it turns out that the person who is feeding you a bunch of numbers can't tell you where they came from. That should be your first hint that you need to be very skeptical about what you are being told.

   Even if your data have an identifiable source, you still want to know what it is. You might have some extra questions to ask about a medical study on the effects of secondhand smoking if you knew it came from a bunch of researchers employed by a tobacco company instead of from, say, a team of research physicians from a major medical school, for example. Or if you knew a study about water safety came from a political interest group that had been lobbying Congress for a ban on pesticides.

   Just because a report comes from a group with a vested interest in its results doesn't guarantee the report is a sham. But you should always be extra skeptical when looking at research generated by people with a political agenda. At the least, they have plenty of incentive NOT to tell you about data they found that contradict their organization's position.

   Which brings us to the next question:

2. **Have the data been peer-reviewed?** Major studies that appear in journals like the New England Journal of Medicine undergo a process called "peer review" before they are published. That means that professionals - doctors, statisticians, etc. - have looked at the study before it was published and concluded that the study's authors pretty much followed the rules of good scientific research and didn't torture their data like a middle ages infidel to make the numbers conform to their conclusions.

   Always ask if research was formally peer reviewed. If it was, you know that the data you'll be looking at are at least minimally reliable.

   And if it wasn't peer-reviewed, ask why. It may be that the research just wasn't interesting to enough people to warrant peer review. Or it could mean that the research had as much chance of standing up to professional scrutiny as a $500 mobile home has of standing up in a tornado.

3. **How were the data collected?** This one is real important to ask, especially if the data were not peer-reviewed. If the data come from a survey, for example, you want to know that the people who responded to the survey were selected at random.

In 1997, the **Orlando Sentinel** released the results of a poll in which more than 90 percent of those people who responded said that Orlando's National Basketball Association team, the Orlando Magic, shouldn't re-sign its center, Shaquille O'Neal, for the amount of money he was asking. The results of that poll were widely reported as evidence that Shaq wasn't wanted in Orlando, and in fact, O'Neal signed with the Los Angeles Lakers a few days later.

Unfortunately for Magic fans, that poll was about as trustworthy as one of those cheesy old "Magic 8 Balls." The survey was a call-in poll where anyone who wanted could call a telephone number at the paper and register his or her vote.

This is what statisticians call a "self-selected sample." For all we know, two or three people who got laid off that morning and were ticked off at the idea of someone earning $100 million to play basketball could have flooded the **Sentinel**'s phone lines, making it appear as though the people of Orlando despised Shaq.

Another problem with data is "cherry-picking." This is the social-science equivalent of gerrymandering, where you draw up a legislative district so that all the people who are going to vote for your candidate are included in your district and everyone else is scattered among a bunch of other districts.

Be on the lookout for cherry-picking, for example, in epidemiological **(a fancy word for the study of disease that sometimes means: "We didn't go out and collect any data ourselves. We just used someone else's data and played 'connect the dots' with them in an attempt to find something interesting.")** studies looking at illnesses in areas surrounding toxic-waste dumps, power lines, high school cafeterias, etc. It is all too easy for a lazy researcher to draw the boundaries of the area he or she is looking at to include several extra cases of the illness in question and exclude many healthy individuals in the same area.

When in doubt, plot the subjects of a study on map and look for yourself to see if the boundaries make sense.

4. **Be skeptical when dealing with comparisons.** Researchers like to do something called a "regression," a process that compares one thing to another to see if they are statistically related. They will call such a relationship a "correlation." Always remember that a correlation DOES NOT mean causation.

   A study might find that an increase in the local birth rate was correlated with the annual migration of storks over the town. This does not mean that the storks brought the babies. Or that the babies brought the storks.

   Statisticians call this sort of thing a "spurious correlation," which is a fancy term for "total coincidence."

   People who want something from others often use regression studies to try to support their cause. They'll say something along the lines of "a study shows that a new police policy that we want led to a 20 percent drop in crime over a 10-year period in (some city)."

   That might be true, but the drop in crime could be due to something other than that new policy. What if, say, the average age of those cities' residents increased significantly over that 10 year period? Since crime is believed to be age-dependent (meaning the more young men you have in an area, the more crime you have), the aging of the population could potentially be the cause of the drop in crime.

   The policy change and the drop in crime might have been correlated. But that does not mean that one caused the other.

5. **Finally, be aware of numbers taken out of context.** Again, data that are "cherry picked" to look interesting might mean something else entirely once it is placed in a different context.

> Consider the following example from Eric Meyer, a professional reporter now working at the University of Illinois:
>
> My personal favorite was a habit we use to have years ago, when I was working in Milwaukee. Whenever it snowed heavily, we'd call the sheriff's office, which was responsible for patrolling the freeways, and ask how many fender-benders had been reported that day. Inevitably, we'd have a lede that said something like, "A fierce winter storm dumped 8 inches of snow on Milwaukee, snarled rush-hour traffic and caused 28 fender-benders on county freeways" -- until one day I dared to ask the sheriff's department how many fender-benders were reported on clear, sunny days. The answer -- 48 -- made me wonder whether in the future we'd run stories saying, "A fierce winter snowstorm prevented 20 fender-benders on county freeways today." There may or may not have been more accidents per mile traveled in the snow, but clearly there were fewer accidents when it snowed than when it did not.

It is easy for people to go into brain-lock when the see a stack of papers loaded with numbers, spreadsheets and graphs. (And some sleazy sources are counting on it.) But your readers are depending upon you to make sense of that data for them.

Use what you've learned on this page to look at data with a more critical attitude. (That's critical, not cynical. There is a great deal of excellent data out there.) The worst thing you can do as a writer is to pass along someone else's word about data without any idea whether that person's worth believing or not.