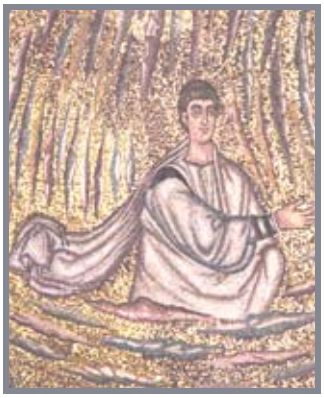
Statistics

Modern nutritional and medical research relies heavily on statistical analysis of data drawn from many different sources. I'll try to take some of the mystery out of this process for you, and explain why you need to keep a grain of salt handy when reading about it. an essay by Alan Yoder



There are three kinds of lies: lies, damned lies, and statistics.

— Benjamin Disraeli

The above quotation is well known and often invoked by those who wish to diss upon a particular statistical finding. Unfortunately, most people do not have enough training in statistics to tell whether a given finding is worthy of interest or not. The good news is that knowledge of a few simple principles will go a long way. I'll try to give you that knowledge here.

Key Concepts

The mathematical basis of statistics is rooted in the study of random ("stochastic") processes. A fundamental element in this study is something called a *random sample.*

A *sample* is a subset of a set of data that one wishes to study and draw some conclusions about, but does not have the wherewithal to study exhaustively. Let's say I have a million marbles and wish to get a rough idea of how many red, blue and green marbles there are. Counting them will take too long. Instead, I get 100 of them, count the red, blue and green ones and multiply by 10,000. My assumption is that the other 10,000 samples in the pile are all the same, plus or minus, as my sample. Statistical theory tells me that if the sample is a *random* sample, my assumption will be a good one. I'm fine with a small error in the result of course, because I don't want to count every marble, but it's good to know the answers I get will be close.

For a sample to be truly random, the items in it must be *independent*. Suppose the million marbles are in a big cement mixer, and first a bunch of red marbles was dumped in, then some blue ones, and the whole thing topped off with green ones. The marbles will be all clumped together by color, and simply scooping up 100 marbles off the top with a scoop will likely give me 100 green marbles. This is not a fair representation of what's in the mixer, because the position of each marble at this point is not independent from the position of its neighbors. To randomize the positions and fix this, I turn on the cement mixer and rotate the drum a few hundred times. Now the marbles are all mixed together, and a scoop off the top should be a random and therefore representative sample.

Another fundamental concept in statistics is *correlation*. Two samples are correlated if there is some evidence of a linear relationship between them. For instance, there is a known high correlation between getting a moderate amount of exercise and having good health. Not everyone who exercises has good health, and not everyone who has good health exercises, but the two things go hand in hand often enough for us to declare them to be correlated.

It's important to understand that the statistics are not saying one thing causes the other, only that they tend to occur together. A great example of the usefulness of this kind of thing is a story from the early days of Walmart's foray into data mining. They found an unexpected correlation between

sales of diapers and sales of six packs of beer. Turns out that when Mom sends Dad out for more diapers, he pretty often picks up a six pack while he's at the store. This little tidbit led Walmart to organize their stores in such a way that when Dad was heading for the counter from the diaper section, he went past the beer. Result: increased sales of both diapers and beer.

There are basically five levels of correlation that you need to know:

(1) No correlation
(2) Statistically insignificant correlation
(3) Statistically significant correlation
(4) High correlation
(5) Perfect correlation

Suppose you are trying to figure out whether people who buy diapers also tend to buy beer at *your* store. You run the numbers and get a "coefficient of correlation," which is a number between zero and one that's usually called *r* and is between zero and one. Things then fall out roughly as follows:

(1) If *r* is near zero, the samples have no discernible linear similarities. There may be very significant non-linear similarities; see the wikipedia article on correlation for a nice illustration of this. But non-linear similarities are not captured by correlation coefficients that medical and nutritional researchers use.
(2) In statistics books there are tables of the *critical values* of *r* . These vary depending on the sample set size you have and the degree of certainty that you want. If your computed value of *r* is less than the critical value, your correlation is statistically not significant and it's probably not worth rearranging your store to put the beer near the diapers.
(3) If your computed value of *r* exceeds the critical value for your sample set size and confidence requirements, on the other hand, the correlation is said to be statistically significant. Consider rearranging your store.
(4) A value of *r* well above the critical value gives fairly deep confidence in the correlation. Is it possible your store is already optimally arranged?
(5) Perfect correlation generally means you've somehow taken the correlation of identical sample sets. Probably you made a mistake somewhere.

One other core concept is the *Law of Large Numbers*. This law states that as an increased number of samples are taken, the statistics for the combined results converge toward the statistics for the actual data very quickly. In a way, the name is misleading, because an astonishingly small number of samples is often enough to guarantee accuracy to near 100%.

Self selection

Most researchers get the correlation calculations right. They will even publish their numbers for *r*, the critical value, sample set size and degree of confidence. They may also publish other numbers such as the standard deviation of the data sets, which is used to compute the correlation. So far, no problem.

The issue is the difficulty of selecting truly random sample sets. Let's go back to the marbles in the mixer. Suppose we know there are 400,000 red marbles, 100,000 blue marbles, and 500,000 green ones. We mix them all up and take our sample of 100, expecting to find 40 red, 10 blue and 50 green marbles, or close to it. Instead there are 80 blue ones with the rest red and green! What went wrong? After some head scratching and poking and prodding, we figure it out: the red and green marbles have lead in the glass; being heavier they sank to the bottom during mixing.

This is an example of a sample set that has *self selected*. The positions of the marbles in the mixer are not independent by color, because the color is perfectly correlated to weight, weight affects how the marbles will mix, and mixing was our method for guaranteeing a random sample. To get a truly random sample, we'll have to find another way.

Self selection in human trials

This is one of the two biggest problems with medical and nutritional studies. It is

notoriously difficult to obtain truly random samples of humans. Humans group themselves in many many different ways: by neighborhood, by race, religion and creed, by sex, health level, age, and education level. Many groupings are subtle to the point of near invisibility. Yet they can have the effect of causing researchers to identify correlations incorrectly. A recent example is the Stanford University study which found that pedometer users walk or run an average of a mile a day over non-users. This was reported uncritically by most of the media as usual. Some suggested, however, that possibly pedometer users are people who are more concerned about their health, and who therefore walk or run more anyway, and that pedometer use was a symptom of that, not a correlation of independent variables.

Correlation vs. causation

In addition to getting the correlation wrong, the study's reporters—and possibly the authors themselves—had a tendency to confuse correlation and causation. The clear implication (gleefully jumped on by pedometer manufacturers and retailers) was "Use a pedometer and you'll walk more!" This statement implies a cause-and-effect relationship between pedometer use and exercise level. The actual correlation that the researchers found does no such thing; it merely states that people who exercised more were more likely to have used pedometers. See the difference?

Data mining

It is not often noted, but probably the majority of nutritional studies are actually what amount to database queries over very large datasets put together by the National Institute of Health and other federally funded programs in academia, non-profit institutes and government departments. Very large datasets have the nice property that self selection is both less likely and less important because of the sheer numbers involved. Less likely, I say. There can be significant exceptions. In the case of census data, for example, illegal immigrants tend to self select out of the data set. In the case of health care data, people who avoid going to doctors and hospitals tend not to be represented. Obese and bulimic people alike tend to self select out of dietary studies on account of embarrassment over their condition.

Often, as in the case of census data, these datasets persist for a decade or more. All dataset gathering techniques have flaws, but the flaws tend to change from time to time. Additionally, new data points are added and others taken away. It is simultaneously amusing and frustrating to those who understand the process to observe the wholesale changes in public health doctrine as the transition from one large and influential dataset to another is made. Thus we have generation-long doctrinal shifts such as the one away from breast feeding to formula and now back again, and the similar one away from butter toward margarine and back again.

Publish or perish

Academic and government researchers are often under pressure to publish findings, which means of course that they have to develop some. This leads to a temptation to publish findings on correlations that barely meet the test for statistical significance. Or that don't even make much sense. It seems that almost any quantitative study can get published nowadays, and statistical studies *are* quantitative, you have to give them that.

It makes more sense to recognize and adjust for this human tendency than to try to change it.

Selective acceptance of results

Given the heightened risks of bad sampling inherent in the analysis of human data, my personal opinion is that only results which exhibit high degrees of correlation should be given much credence. An alternative approach might be to modify the "critical values" tables to make allowance for these risks when dealing with human data.

There is no current movement in this direction that I'm aware of, either in academia, the medical and nutritional establishment or in the press. It is left for us lay

people to make the required adjustments in what we believe and what we ignore.

Reporting

There are few newspaper reporters working for the non-technical press who understand these issues. Most reporting of nutritional and medical research, and especially of research based on statistical studies, is uncritical to the point of vacuity. Efforts to add value to the story usually end up either distorting the results or implying causation when none was indicated, as in the pedometer story.

Again, caveat emptor. It's up to us to sort this out.

Statistical research considered harmful

Statistics itself is a noble branch of mathematics with an extremely rigorous and clear foundation. See my essay on Science for some information about this. So in a sense Disraeli was wrong.

But in any case, the problems I've listed in the last several sections are so pervasive that stories about nutritional research in the daily press are best left unread unless there are links to deeper sources. As reported they are almost always meaningless or downright misleading. My advice is to ignore them.

How to proceed

Instead of worrying about nutrients, think (joyfully!) about food. Eat simply and well, using wholesome ingredients. Eat things that your ancestors 100 years ago would have recognized as food. My cookbook is a good source for this. I have heard that another good work on this subject is *In Defense of Food* by Michael Pollan. I have not read this yet, but plan to.