Viewpoint: Why so many scientific studies are flawed and poorly understood

Should we believe the *USA Today* headline, “Drinking four cups of coffee daily lowers risk of death”? And what should we make of, “Mouthwash May Trigger Diabetes. . .”? Should we really eat more, not less, fat? These sorts of conclusions, supposedly from “scientific studies,” seem to vary from month to month, leading to ever-shifting “expert” recommendations. However, most of their admonitions are based on dubious “research” that lacks a valid scientific basis and should be relegated to the realm of folklore and anecdotes.

Flawed, misleading research is costly to society because much of it is the result of poorly spent government funding, and it often gives rise to unwise regulation. One remedy would be greater statistical literacy that would enable the public—and their elected leaders—to reject “junk” science.

Statistics is a mathematical tool used in many scientific disciplines to analyze data. It is intended to provide a result that will reveal something about the data that otherwise is not obvious, which we will refer to as a “finding” or a “claim.” Before undertaking an analysis, a researcher formulates a hypothesis—which is his best guess for what he expects to happen.

A “p-value” is a term used in statistics to indicate whether the finding confirms the result that the researcher was expecting. An essential part of this process is that *before* undertaking the analysis, the researcher must formulate a hypothesis that he expects the analysis would tend to prove or disprove based on the p-value. The lower the p-value, the greater the confidence that the finding is valid.

Usually a “strawman” hypothesis is advanced, for example that treatments A and B are equally effective.The two treatments are then compared, and any p-value less than 0.05 ($p<.05$) is, by convention, usually considered “statistically significant” and tends to *disprove* the strawman hypothesis that the effects of the treatments are the same. The alternative hypothesis, A is different from B (for example, aspirin is better than a sugar pill, to relieve a headache) is now accepted.

However, and this is a key point—a p-value less than 0.05 ($p<.05$) can occur by chance, which is known as a false positive. The standard scientific approach to identifying a false positive is to attempt to replicate the possibly false positive result. If the original results don’t replicate, it is assumed that they were false—and we’re left with the original “strawman” hypothesis that there is no difference between A and B.

But things can get complicated, because the p-value analysis can be manipulated so that it *appears* to support a false claim. For example, a scientist can look at a lot of questions, which is known as “data dredging,” and formulate a hypothesis *after* the analysis is done, which is known as HARKing, H ypothesis After the Result is Known. Together these violate the fundamental scientific principle that a scientist must *start* with a hypothesis, not concoct one after the data set has undergone analysis.

A simple coin-toss example illustrates the point. Say a scientist is analyzing 61 flips of a coin, and at some point there are five successive heads in a row. Seeing this result, the scientist formulates a hypothesis that this result, unexpected taken in isolation, seems to prove the coin is unfair. The perception of unfairness of the coin can be bolstered by not revealing that there were 56 other tosses of the coin in the sequence.

The claim is, of course, a false positive because on the next set of 61 coin tosses it is unlikely that there would be five successive heads *at the same place* in the new sequence. In Table 1 we present ten 61-toss sequences. The sequences were computer generated using a fair 50:50 coin. We have marked where there are runs of five or more heads one after the other.

In all but three of the sequences, there is a run of at least five heads. Thus, a sequence of five heads has a probability of 0.5^5=0.03125 (i.e., less than 0.05) of occurring. Note that there are 57 opportunities in a sequence of 61 tosses for five consecutive heads to occur. We can conclude that although a sequence of five consecutive heads is relatively rare taken alone, it is not rare to see at least one sequence of five heads in 61 tosses of a coin.

image

Image not found or type unknown

Table 1. Given are 10 sequences of 61 coin flips each, 1= heads and 0=tails. Note that 5 consecutive heads occur (0.5)5=0.03125 rarely, ~3% of the time, but with 61 flips, are found in 7 of the 10 sequences, with 61 flips. In none of the 10 sequences do the runs of 5 heads appear at the same place in the sequence.

Now, let us consider a food consumption experiment. We simulate the results of a food frequency questionnaire, or FFQ, with 61 different foods and their possible health effects. In such an experiment, a very large number of people are asked how much of these 61 foods they typically eat. Later, the people answer a heath questionnaire containing questions about whether they have experienced high blood pressure, gastric reflux, a history of pancreatic cancer, etc.

The first such study did, in fact, inquire about 61 foods. There were many health effects collected in the later survey. For this simulation, in order to illustrate the fallacy of such studies, we will have 10 health effects, which are numbered: HE 1…HE 10.

The kind of question of interest to investigators might be, “Does eating an orange every day reduce cholesterol?” Thus, there are 61 x 10 questions at issue and they can be arranged in a 61 by 10 table (Table 2). It is usual to declare “statistical significance” if the p-value for any of these 610 questions has a p-value <0.05, and we can use a computer to simulate statistical significance.

In Table 2, we have placed a “1” where the simulated p-value was less than or equal to 0.05 and a “0” in any cell where the simulated p-value was greater than 0.05. Each column in the table represents a

separate health effect. Note that in this simulation each column (health effect) has a significant p-value. What are the chances of at least one statistically significant—but not real— correlation in a 61-food experiment with only one health effect examined? It turns out that the probability is very high—about 0.95, where 1.00 means that it happens every time.

Of course, looking at more health effects increases the chances of a statistically significant result somewhere in the study. With 61 foods and 10 health effects the chance of a nominally significant result by chance alonea statistically false positive, “fake” result—is essentially assured. We appear to be viewing a false-positive-generating machine.

31	1	0	0	0	1	0	1	1	0	0
32	1	1	1	0	1	1	1	1	1	0
33	0	1	0	1	1	0	0	0	1	0
34	0	0	1	0	0	1	1	1	1	1
35	1	1	1	0	1	1	0	0	1	0
36	0	0	0	1	1	0	0	1	0	0
37	1	0	0	1	0	0	1	0	1	1
38	1	0	1	0	1	0	0	0	1	0
39	0	0	1	1	1	0	0	0	1	1
40	1	1	1	1	1	0	1	1	0	0
41	0	0	1	1	1	0	1	1	0	1
42	1	0	0	0	1	1	0	1	0	0
43	1	1	0	1	0	0	1	1	0	0
44	1	0	1	0	1	0	1	1	0	0
45	1	0	1	1	1	1	0	1	1	0
46	1	0	0	0	1	1	0	0	0	1
47	0	0	0	1	1	0	0	1	1	0
48	1	0	0	0	0	1	0	1	1	0
49	1	1	1	1	0	0	0	1	0	0
50	0	1	1	1	1	0	0	1	1	0
51	1	0	0	0	0	0	0	1	0	0
52	1	1	0	0	0	0	0	1	1	1
53	1	0	1	1	0	1	0	0	1	0
54	0	0	0	1	0	1	0	0	1	1
55	0	0	0	1	1	1	1	0	1	1
56	0	0	0	0	0	0	1	1	1	0
57	0	1	1	1	1	1	1	1	1	1
58	0	1	0	1	0	0	0	1	1	1
59	0	1	0	0	1	0	1	0	0	1
60	0	0	0	1	1	0	1	0	0	0
61	1	0	1	0	1	0	1	1	0	1

Table 2. There are 61 foods, rows, and 10 health effects, columns. A “1” indicates statistical significance, p<0.05, and a “0” indicates no nominal statistical effect. Each “1” is a statistical false positive. For each “1” a paper could be written about a finding that would not be expected to replicate.

But surely the difference between asking one question (one food and one health effect) and asking 610 questions is well-known to researchers. Well, yes, but asking lots of questions and doing weak statistical testing is part of what is wrong with the self-reinforcing publish/grants business model. Just ask a lot of questions, get false-positives, and make a plausible story for the food causing a health effect with a p-value less than 0.05: HARKing.

The first published Food Frequency Questionnaire (FFQ) came out of the Harvard School of Public Health and had 61 questions. For any health effect, asking 61 questions gives about a 95% chance of getting a

statistically significant result – which may or not be "real"— for each health effect. Thus, the critical point is: *Beware of any study that asks too many questions!*

But for many FFQs, 61 questions were not enough. More recent versions ask even more. A paper in 2008 used a FFQ with 131 questions, which were asked at two different time points, giving a total of 262 questions. They reported an association between women eating breakfast cereal and increased odds of having a boy baby. (For the record, the sex of a zygote is determined by whether the male's sperm contributes an X or Y chromosome.) A U.S. government survey uses a FFQ with 139 questions, and a recent paper that appeared in the journal *Heart* and used a FFQ with 192 food questions found a decrease in atrial fibrillation associated with chocolate consumption.

"Data dredging" and HARKing that yields false-positive results can also be applied to laboratory animal experiments, as explained here by Dr. Josh Bloom, a chemist at the American Council on Science and Health. Those phenomena apply as well to clinical studies. Consider this caveat from an article in JAMA, which critiqued an article about a medical device to prevent stroke during the replacement of the aortic valve via a catheter:

Statistically comparing a large number of outcomes using the usual significance threshold of .05 is likely to be misleading because there is a high risk of falsely concluding that a significant effect is present when none exists. If 17 comparisons are made when there is no true treatment effect, each comparison has a 5% chance of falsely concluding that an observed difference exists, leading to a 58% chance of falsely concluding at least 1 difference exists.

Spurious FFQ studies are published constantly. The inventor of the FFQ has to his credit (?) more than 1,700 papers. The original FFQ paper is cited over 3,300 times. It appears that virtually none of the researchers using FFQs correct their analysis for the statistical phenomena discussed here, and the authors of FFQ papers are remarkably creative in providing plausible rationales for the "associations" they discover—in other words, HARKing.

This situation creates a kind of self-licking ice cream cone: Researchers have been thriving by churning out this dubious research since the early 1990s, and inasmuch as most of the work on Food Frequency Questionnaires is government funded—by the National Cancer Institute, among other federal entities—it's ripping off taxpayers as well as misleading them. Curiously, editors and peer-reviewers of research articles have not recognized and ended this statistical malpractice, so it will fall to government funding agencies to cut off support for studies with flawed design, and to universities to stop rewarding the publication of bad research. We are not optimistic.

Dr. S. Stanley Young is a statistician who has worked at pharmaceutical companies and the National Institute of Statistical Sciences on questions of applied statistics. He is a member of the EPA's Clean Air Science Advisory Committee.

Henry I. Miller, a physician and molecular biologist, is the Robert Wesson Fellow in Scientific Philosophy and Public Policy at Stanford University's Hoover Institution. He was the founding director of the FDA's Office of Biotechnology. Follow him on Twitter @henryimiller.