# Odds Are, It's Wrong

## The Misuse of Math in Science, Medicine and the Media

Tom Siegfried

Alfred and Julia Hill Lecture
University of Tennessee
March 31, 2009

# Odds Are, It's Wrong:
# The Misuse of Math in Science, Medicine and the Media

**Abstract:** A lot of science is not properly done, and even when properly done it is often misreported or misinterpreted — sometimes by the media, sometimes by scientists themselves. At the root of this problem are conventions for drawing conclusions based on inferences from statistical analysis, using statistical methods that are, in essence, bogus. As a consequence, the modern system of media coverage of science news from journals has created conditions that almost guarantee that the scientific studies most likely to be reported in the mass media are also the most likely to be wrong. These problems are especially egregious in medicine, most notably in clinical trials — touted as the "gold standard" for medical research but actually deeply flawed.

A few years ago, a paper appeared in the journal *PLoS Medicine* with a rather provocative title: "Why Most Published Research Claims are False." It was written by John P.A. Ioannidis, an epidemiologist with joint appointments at a medical school in Greece and at Tufts University School of Medicine in Boston. At first glance it seemed like a headline from that satirical newspaper the *Onion*. But Ioannidis was clearly serious. I called Donald Berry, a respected biostatistician at the University of Texas M.D. Anderson Cancer Center in Houston, to ask what he thought of the Ioannidis paper. "He took a somewhat extreme view," Berry told me. "But the bottom line is, he's right."

Essay

**Why Most Published Research Findings Are False**

John P. A. Ioannidis

**Summary**

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for

factors that influence this problem and some corollaries thereof.

**Modeling the Framework for False Positive Findings**

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

**It can be proven that most claimed research findings are false.**

In his paper, Ioannidis outlined a wide spectrum of statistical disorders that plague research in scientific fields ranging from social psychology to molecular biology. He contended that the standard methods of applying math to scientific data rendered any single scientific study's conclusion likely to be incorrect.

"There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims," Ioannidis declared.

For science journalists, who have all been taught that the first thing you ask about a research result is whether it is "statistically significant," this revelation would come as something of a shock. But it shouldn't. Weaknesses in the standard statistical approach have been described in the scientific literature for decades. In fact, if you

believe what you read in the scientific literature, you shouldn't believe what you read in the scientific literature.

This problem has widespread ramifications, for science and for medicine and public health and all sorts of public policy. Problems with statistical methodology are behind a lot of the problems in medical policy making today, for example, and they explain why so many media reports are often confusing — why a coffee is bad in one study but OK in the next, why certain therapeutic drugs were approved but then turned out to be deadly. Bad statistics are behind some very bad medical decisions, as one recent *Onion* headline made very clear.

the ONION

AMERICA'S FINEST NEWS SOURCE · ONION.COM

WEATHER

FDA APPROVES SALMONELLA

Pathogen 'Now Part Of A Well-Balanced Diet'

WASHINGTON—Calling it "perfectly safe for the most part," and "not nearly as destructive or fatal as you'd think," the Food and Drug Administration approved the enterobacteria salmonella for human consumption this week.

The federal agency, which has struggled in recent years to contain the food-borne pathogen, and repeatedly failed to prevent tainted products from reaching store shelves, announced Monday that salmonella was now completely okay for all Americans to enjoy.

"Rigorous testing has shown that salmonella is...fine," FDA director of food safety Stephen Sundlof said. "In fact, our research indicates that

see FDA, page 7

FDA director Stephen Sundlof okays the bacteria for eating, drinking, and applying directly to the skin.

Now, I'm not saying that *all* science is wrong or that scientific research never produces reliable knowledge. It does. The last impression I want to leave is that science is worthless and should be casually disregarded. But arriving at reliable knowledge is more complicated than usually represented. And that complexity is almost always off the radar for most journalists. And so the science you read about in the news is especially susceptible to being misleading, precisely because of these statistical issues. In fact, the criteria used by journalists, including science journalists, to evaluate newsworthiness almost guarantee that the research results most likely to be reported in the media are also the results most likely to be false.

This situation poses a very serious problem for science journalism, a problem that is not widely recognized. The presumption seems to be that if scientists publish a finding in a peer-reviewed journal then it should be reported as though it were an addition to accepted knowledge, if it otherwise meets the criteria for news. But the criteria for news are almost exactly the criteria you would choose to select false findings, not true ones. That's because false studies are very likely to exhibit one or more of three most common features that science journalists seize on when deciding what stories to write. The first of these is the very idea of "first." If a study is the "first" to find or report something, that makes it more likely to be news. It is also likely to be wrong. A second news criterion is reader interest, which translates into particular journalistic interest in hot research fields, where even incremental "advances" are likely to become news. Such "hot field" advances are, it turns out, also among the most likely to be wrong. The third flag that captures journalists' attention is "contrary to previous belief." A report that contradicts conventional wisdom is likely to make news. It is also likely to be wrong.

If that is all obvious to you, I hope the tickets to this talk were all free. But I seriously do hope that you now wonder, How can this be? That's what I'm going to try to explain.

My plan is to briefly outline the basic ideas of using statistics to analyze experimental data and draw inferences or conclusions, and then to describe some ways that these numbers are commonly misreported and misinterpreted. Then I'll discuss some examples that show why these methods, while possibly sound in theory, don't work very well in practice, as in studies linking genes to disease or clinical trials for testing new medicines. And then I'll say something about why the situation isn't completely hopeless, and how science and science journalism can deal with the deficiencies in currently common practices.

## Statistics Basics

First of all, let me emphasize that statistics is an extremely sophisticated and complex branch of mathematics. Nothing I say should be taken as trivializing it, and you should be warned up front that there's a lot more to statistics than what I will present. I'll be talking about statistics in simplified terms for purposes of illustration, so please promise to remember that everything is actually much more complicated than I say it is.

Fortunately, there is a good everyday example of statistics in action that can be used to illustrate the principles involved: public opinion polls. When you read that a poll result has a margin of error of 3 percentage points, that means that in theory you'd get a result in that range 95 percent of the time. But where does that margin of error come from? It is based on the number of people that the pollsters sample, and is computed with the help of some relatively simple math. Here it is:

$$\pm 1.96\sqrt{\frac{(100-P)P}{n}}\times\frac{(N-n)}{N-1}$$

This is the formula that pollsters use to compute a 95 percent-confidence margin of error. (P is the percentage of people in the sample who give the answer you're interested in; n is the number of people you poll, N is the number of people in the entire population that you are sampling from.)

This equation shows that you don't need to poll everybody to get a good idea of what most people think. If you sample some of the people, chosen at random, the answers they give will pretty much reflect the answers you'd get from asking everybody — IF you have a big enough sample and if it is truly randomly chosen from among all the people in the population. When that's the case you can use this simpler formula.

$$\pm 1.96\sqrt{\frac{(100-P)P}{n}}$$

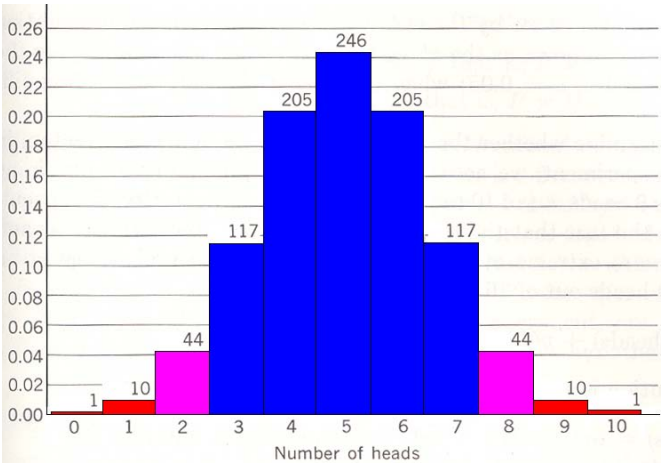
It tells you what range of answers you'd get if you sampled the population over and over again — the results would be within that range 95 percent of the time. (In theory, of course. In real life all sorts of practical problems conspire to make the margin of error bigger.)

Now the key thing here is — you can do the calculation yourself or you can just believe me — is that it doesn't really matter how big the population is. What matters is just the size of the sample, pretty much regardless of how big a population you took it from.

We know the idea behind this math is correct because when you apply it to simple, easily controlled situations, it gives the right answers. Such a simple controlled situation would be flipping coins. If you take a penny and toss it 10 times and count the number of heads, you'll probably get more than one head but fewer than 10. And if you keep doing these 10-flip sessions over and over again, you'll find that most of the time you get from 3 to 7 heads. Over a long time, on average, you will get roughly half heads and half tails, unless there is something fishy about your penny.

The neat thing about this math is it will tell you not only that you get roughly half heads and half tails in the long run, but also how often you should expect to get any particular number of heads. It tells you what percentage of the time that you'll get 0 or 1 or 4 or 9. So you can test whether your penny is fishy by doing this experiment, and the result will tell you how confident you can be that your penny is fair.

Here's the chart. If you do the 10 flips and get fewer than 3 heads, that would be a result you should expect to happen only 5.5 percent of the time (summing the chances of getting 0, 1, or 2).




As I mentioned, in real science the math gets much more complicated. But the core ideas behind the math applied to polls and pennies still apply when doing statistical inference in almost any kind of study done using samples, which is just about anything.

## P values

Until less than a century ago, scientists didn't really know much about this stuff or apply it systematically to experimental comparisons (although it had long been used in estimating the size of random measurement errors). The modern era began with the British mathematician Ronald Fisher, who devised something called the P value approach.

P values, textbooks will tell you, are a measure of the likelihood that experimental results indicating an apparent effect are due to chance. A P value of 0.05 means that the measured magnitude of an effect would be expected to occur by chance 5 percent of the time (or 1 time in 20). A P value of 0.01 means that the measured magnitude of an effect would be expected to occur by chance 1 percent of the time (1 time in 100).

Fisher believed that the right way to do science was to set up experiments that could be analyzed with statistics by computing a P value. Suppose you wanted to see if fertilizer made corn ears bigger. (Fisher actually looked at crop yields, but I think it's easier to illustrate the idea with corn ears.) You would plant some corn, some fertilized and some not, and then measure the size of the ears for each group.

Because corn ear size varies anyway, no doubt that average ear size of the fertilized corn would come out a little different from the average for unfertilized corn. The question is whether that difference is really because of the fertilizer, or just the usual random variations. Fisher proposed to address that question this way: First, assume that the fertilizer had no influence on ear size. (That is what he called the "null hypothesis.") Then, using the formulas of statistics, calculate how likely it was that you would see a difference as large or larger than what you actually measured. If you'd expect to see a difference that big fewer than one time in 20 trials, Fisher said you should reject the null hypothesis, and conclude that fertilizer did increase the size of the corn ears.

Just in case that went by too quickly, I'll summarize it again:

- Null hypothesis: Fertilizer does not work.

- Compare: Average ear size+variation for fertilized corn
  Average ear size+variation for unfertilized corn

- Apply formula to compute probability of seeing a difference that big or bigger

- If that probability is less than 5 percent, then reject the null hypothesis, and conclude that fertilizer *does* work.

Fisher realized, of course, that the no-effect hypothesis might actually still be right — that maybe your experimental results were just an unusual fluke— but nevertheless if you got a result to be expected less than 5 percent of the time you ought to bet that the no-effect hypothesis was wrong, and conclude that fertilizer works. And that's what most scientists still believe today.

While Fisher's approach became very influential, it didn't satisfy everybody. In particular, Jerzy Neyman and Egon Pearson came along and said they had a better idea — not to test a null hypothesis, but to test two competing hypotheses. Their approach, by a more elaborate mathematical route, also produced a P value. Their P value refers to the likelihood of a false positive (concluding an effect is real when it actually isn't) or false negative (concluding that there is no effect when in fact there actually is).

All this was going on in the 1920s and '30s. What eventually emerged was a hybrid mix of these two mutually inconsistent approaches, which has left the interpretation of standard textbooks statistics terribly muddled. In fact, this whole system, while widely used, really does not work very well.

In actual fact — and Fisher, Neyman and Pearson realized this — when you do statistical analyses of this sort you can't actually be sure of anything. If you flip a penny 10 times and get only one head, maybe the penny is rigged, or maybe it was just one of those rare events that happen. Maybe fertilizer works, Fisher would say, or maybe you just witnessed an unusual set of data. All you can say that the result is or is not "statistically significant," based on whatever degree of probability you'd like to arbitrarily select—usually 5 percent (but much lower in some fields). In other words, if the chance of getting certain data is less than 5 percent, that result is considered to be statistically significant (at the 5 percent level).

What I'm here to say is that there are huge problems with how such results are interpreted and presented, especially by journalists in news accounts but also often by scientists themselves.

## Statistical significance

These issues all revolve around a high degree of confusion about the meaning of the phrase "statistical significance." Basically, there are three main points to make:

1. A statistically significant effect is not the same thing as a significant effect.

2. Lack of statistical significance does not mean there is no effect.

3. Statistical significance says nothing about the likelihood that an effect is true.

First, I cannot emphasize too much that statistical significance is not the same thing as significance. Often you'll read in a study that some pollutant significantly increased the risk of cancer, or something like that. But most of that time that's just some journalist's shorthand translation of "statistically" significant.

It very well may be that the effect was also significant in the ordinary usage of the word, but not necessarily. Basically there are two ways for an effect to be statistically significant — either it is a very big effect, or the study had a very big sample. If you have a huge sample size, very small effects will show up as statistically significant. A new drug may be *statistically* significantly better than an old drug, but it that might mean that for every thousand people you treat, you get 1 or 2 additional cures. That's not clinically significant. And some studies claim a chemical causes a "significantly increased risk of cancer" when it is just statistically significant, but only a tiny absolute increase.

Here's a simplified example. Suppose you tested a new a cancer drug on 20,000 people and compared survival rates to another group of 20,000 people on standard treatment. In the standard group, 15,000 people die. In the new drug group, 14,800 people die.

That difference is very small, clinically meaningless, but the math says it is statistically significant. Now suppose you tested another new drug on 20 people, comparing it to 20 people on standard treatment. In the standard treatment group, 15

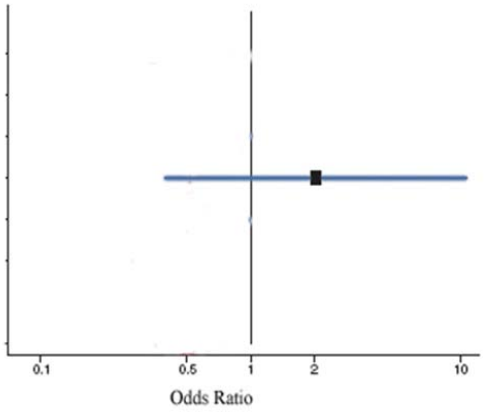
people die. In the test group, only 10 people die. That doubles the number of survivors. But it is not statistically significant, because the sample size is so small.

This distinction between statistical and actual significance is almost never made in news reports, and it is frequently unclear even in scientific reports.

This example also illustrates my second point, which is that lack of statistical significance does not necessarily mean that there is no effect. It may just be that the study as designed was not capable of detecting a real effect. Typically that is because the study is too small, but it could be for other reasons having to do with poor study design. In any case, it is wrong to conclude that there is no effect simply on the basis of lack of statistical significance. A drug is not proven "safe" if the increase in deaths attributed to is not statistically significant, or a new treatment is not worthless just because the number of lives it saves does not meet the statistical significance threshold.

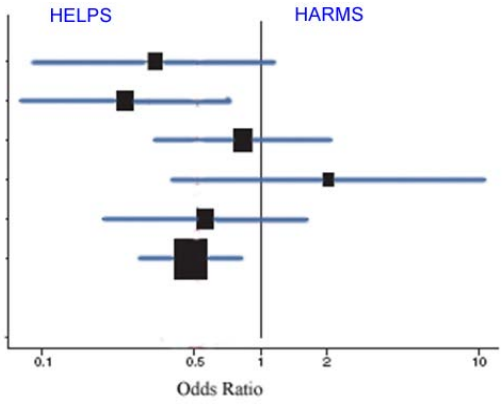
To illustrate that point it helps to mention the idea of confidence intervals. Confidence intervals are just like the margin of error in a poll; they describe the range of results that would be expected within the limits of a given P value. Usually the confidence interval is computed at the 95 percent level, which means that the results within that interval would have a P value of 5 percent or less.

In my cancer drug example, the odds of surviving was doubled. But because the sample was small, the confidence interval — the margin of error — was huge. The range of survival odds at the 95 percent confidence level was 0.8 to 11.5. So lack of statistical significance does not mean the drug doesn't work — it might be 10 times better than the old drug. You just don't have strong enough evidence to conclude with confidence that it works.




Here's a real-life example. One study of this drug showed no statistical significance in effectiveness. But other studies showed that it really did work.

This works the other way around, too. If a food additive is causing cancer, a study might measure an increased risk that is not statistically significant. But the confidence interval might indicate that it still could be dangerous, so you can't conclude that the additive has been shown to be safe simply because there is no statistical significance in the results.




## Rejecting the null hypothesis

The good news is that the confusion between real significance and statistical significance is widely recognized and easily

corrected. Another problem with the standard statistical approach is not so simple and is a continuing source of misunderstanding. And that involves how to interpret the P value that tells you whether you have a statistically significant result.

Say you get a result that is likely to occur less than 5 percent of the time if the null (no-effect) hypothesis is correct. The common temptation is to conclude that therefore it is 95 percent certain that the no-effect hypothesis is wrong, and that you can conclude with 95 percent confidence that the effect is real.

That reasoning is very common, even among scientists. But it is wrong, wrong, wrong, 95 times wrong.

For one thing, that likelihood was based on the assumption that the no-effect hypothesis is correct. If it isn't, then you can no longer say that the likelihood of getting that data was only 5 percent.

But apart from that, it's just not a proper logical conclusion. It's a subtle point, so I've tried to come up with an example to illustrate the problem.

Here's the usual logic:

- If the no-effect hypothesis is true, I am unlikely to get this result.
- I got this result.
- Therefore the no-effect hypothesis is probably wrong, so there probably is an effect.

But now look at the same logic in a different illustration:

- If it is winter, I am unlikely to be swimming.
- I am swimming.
- Therefore, it is unlikely to be winter.

That conclusion is not sound. Its validity depends on other information not included in the analysis. Suppose, for instance, that I am in Hawaii. The fact that I am swimming, then, does not imply that it is not winter. It may just so happen that I spend four days each winter in Hawaii (about 5 percent of the winter). And so it is true that if it is winter I am unlikely to be swimming, as I'm only in Hawaii 5 percent of the time in the winter. Nevertheless, the fact that I am swimming is more likely evidence that I am in Hawaii than it is evidence that it isn't winter.

The point here is that statistical inference based on the results of an experiment tell you *nothing* on its own about how likely it is that the effect is real or not. To make a conclusion like that, you need to know other things.

In this example, for instance, you would need to know not only how much time I spend in Hawaii, but also how much time I swim in the summer! If I swim every day in the summer and only a few days in the winter, the fact that I am swimming does indeed suggest that it's probably not winter. But if I *never* swim in the summer and *only* swim a few days each winter in Hawaii, then the fact that I am swimming suggests that it is probably winter — even though I truly am unlikely to be swimming in the winter (as I spend most of the winter in Washington).

Similar considerations apply to science, in much more complicated and less easy to see ways. It's not that nobody realizes this. The scientific literature is full of papers pointing out such flaws of interpretation and other problems with the application of statistical methods. That's especially the case in medicine, but the problem applies across the whole spectrum of the sciences. It's exactly the problem that Ioannidis was calling attention to. But many others have made similar points.

Some examples:

> "… despite the awesome pre-eminence this method has attained in our experimental journals and textbooks of applied statistics, it is based upon a fundamental misunderstanding of the nature of rational inference, and is seldom if ever appropriate to the aims of scientific research."
>
> — William Rozeboom, *Psychological Bulletin*, 1960

> "Many biologists remain confused about appropriate statistical techniques, or unaware of the limitations of the techniques that they use."
>
> — Philip A. Stephens et al, 2007

> "The methods of statistical inference in current use … have contributed to a widespread misperception ... that absent any consideration of biological plausibility and prior evidence, statistical methods can provide a number that by itself reflects a probability of reaching erroneous conclusions. This belief has damaged the quality of scientific reasoning and discourse...."
>
> — Steven Goodman, 1999

> "Too many reports of associations between genetic variants and common cancer sites and other complex diseases are false positives. A major reason for this unfortunate situation is the strategy of declaring statistical significance based on a *P* value alone...." — Sholom Wacholder et al, 2004

> "Many investigators do not know what our most cherished, and ubiquitous, research desideratum — 'statistical significance' — really means. This . . . signals an educational failure of the first order."
>
> — Raymond Hubbard and J. Scott Armstrong, 2006

> What used to be called judgment is now called prejudice, and what used to be called prejudice is now called a null hypothesis. . . . [I]t is dangerous nonsense (dressed up as the 'scientific method') and will cause much trouble before it is widely appreciated as such. — A. W. F. Edwards, 1972

## Misleading media

So after assessing all the evidence and the critiques of statistical methodology that experts have articulated, it is quite clear that the provocative title of Ioannidis' paper was actually right on target. And so therefore much of what you read in the media about scientific results is wrong.

Now, it isn't exactly a news bulletin that the media often get science wrong. But I'm saying something much more striking than that. Even when the journalists get the story right, what you read is likely to be wrong. In other words, even if the reporter quotes the scientist correctly and in context, faithfully presents the conclusions of the scientific paper and represents the meaning of the findings for science just as the scientists do, odds still are that it's wrong.

That would be the case, if we accept Ioannidis' analysis, simply because most scientific papers are wrong to begin with. But with what you read in the news, it's even worse. Because, as I mentioned earlier, the qualities of a scientific paper that make it newsworthy are precisely those that make it even more likely to be wrong.

As I mentioned before, journalists are eager to write about a paper when it's the first report of such a finding, or when it's a development of any sort in a hot research field (and therefore of wide interest), and when it contradicts previous belief. These criteria for news are a recipe for reporting the results most likely to be bogus.

First of all, consider first reports. Although there is some disagreement on this point in the literature, there are good arguments that first reports are the most likely to be wrong. In any case it seems clear that first reports are likely to overstate the magnitude of an effect if there is one. And it may be that first reports are likely to be wrong more in some fields than others.

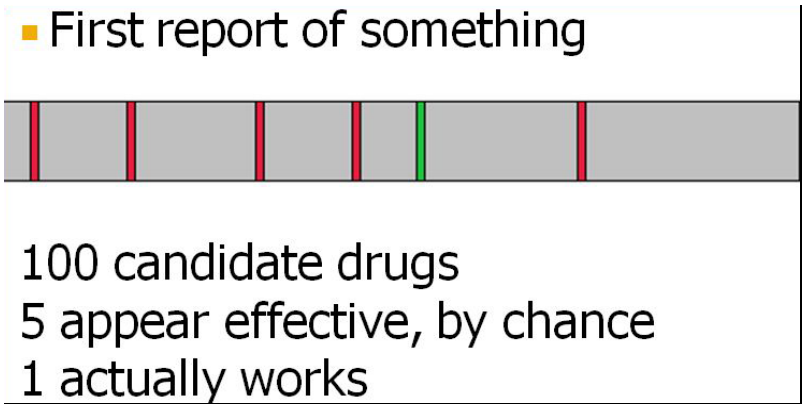
Genetic association studies are one example. One observer writes that "…the initial published report often overestimates the true genetic effect size ... indicating that results from an initial study must be viewed with great caution and may merely be the result of chance, bias, flaws in analysis/interpretation, or publication practices" (Sullivan 2007).

To see why that is, let's suppose, for example, that a new disease arises with no known cure. We'll call it twittering.

Scientists in this simple example world think that some drug already in their arsenal could cure twittering. And their arsenal has 100 drugs. Let's say one of these drugs actually does cure twittering, the other 99 are worthless. But scientists don't know which one will work, so they start testing all 100.

Using that P value of 5 percent as the criterion for statistical significance, we should expect just by chance that five of the drugs will appear to work, even if none of them do. So if only one drug actually works, chances are 5 out of 6 that the first report

of a drug that works will be bogus. This is the same as saying that if you take one of a pair of dice, you'd have to get a 6 on the very first roll for the first report to be correct. Yet it is the first report of a drug to cure twittering that would certainly get the news coverage.




Now when do news reporters tend to write more than just the first report of a result? When it is in a hot research field, like cloning or cancer research or genetics. In hot research fields, lots of labs around the world are all racing to do something or discover something or find evidence of something. So if the race is on to cure twittering, maybe 1,000 labs will be on the job. Over the course of a year maybe 50 papers might be published reporting statistically significant results about drugs or therapies or whatever that reduce death rate from twittering. Of course, the other 950 labs didn't get any results worth publishing. Those 50 papers represent 1 lab in 20 — exactly the 5 percent of the time that the fluke data would expect to appear significant when you set your standard for rejecting the null hypothesis at P = 0.05. News reports on all those stories might give the impression that there are lots of cures for twittering, when there is no good reason to believe any of those results.

Finally, another sure way to get a reporter to write about a study is for the press release announcing it to start out by saying "contrary to previous scientific belief . . . ."

But here again, these are precisely the results that are least likely to stand up.

Presumably (although of course, not always), previous scientific belief is based on previous scientific data. If new data doesn't correspond to the old data, it is most likely that the new data is the statistical outlier. Remember the illustration of confidence intervals, when one study contradicted all the others. Had it been the last study done, it still would have been wrong. There is usually no reason to believe that one new study is right and all previous studies are wrong (although there are cases when that might not be right — if the new data comes from a technologically advanced instrument with higher resolution or something). But ordinarily, contrary to previous belief should be a warning flag that the result being reported is likely to be wrong. Instead it is usually a green light to go with the story.

### Real life errors

So far the examples I've given are simplified and artificial. But there are plenty of real life examples.

Take studies linking genes to disease, for instance. You couldn't count all the news stories over the last 25 years or so that have reported something like "Gene Y is

(statistically) linked to Disease X." Whether those reports hold up might make a good test of whether these statistical methods really work.

Not long ago some researchers did that test. They identified all the papers they could find linking various genes to acute coronary syndrome, a cluster of heart problems. The researchers found 85 genetic variants in 70 different genes that had been linked to increased susceptibility to that syndrome, ACS. Next they identified 811 patients that had been diagnosed with ACS over a period of time at university-affiliated hospitals in Kansas City. If any of those 85 genetic variants are in fact linked to ACS, they ought to show up more in that group of patients than in a control group of people without ACS. So the researchers found 650 controls (matched for sex and age) and did genetic testing on both groups.

Guess how many of the 85 suspect gene variants turned up more often in the people with ACS? Zero. None. Nada. Well, actually one did, but that would be expected just by chance.

"Our null results provide no support for the hypothesis that any of the 85 genetic variants tested is a susceptibility factor for ACS," the researchers reported in the *Journal of the American Medical Association*.

Of course, the search for disease genes grows ever more sophisticated. Nowadays a popular method of testing for genetic links to disease is the use of a microarray, which is basically a computer chip coated with molecules that can sense the presence of chemicals indicating the amount of activity of various genes. Researchers use microarrays to compare gene activity in healthy cells with gene activity in sick cells. It's a way to see which genes are working abnormally in a particular disease.

The problem is that in most diseases, many, many genes are behaving abnormally. And a microarray can test more than 20,000 genes at once. So that raises the immediate issue of how to tell whether a gene is really abnormal or that its activity is just varying within normal limits. You want to know if there is a statistically significant difference, in other words. But if your test for statistical significance is at the 5 percent level, and you test 20,000 genes, then 1,000 genes are likely to appear to be abnormal just by chance!

Setting a higher threshold of statistical significance will eliminate some of those flukes, but only at the cost of also eliminating some truly changed genes from the list as well.

That's a particularly bad problem if you're studying a metabolic disease such as diabetes, when a large number of genes show altered activity but the change in activity is small. When you apply standard statistical approaches, many of these changes fail to reach statistical significance, because the standard tests are weighted to weed out the false positives. Real changes are chalked up as mere fluctuations. So even though hundreds of genes might actually be altered in diabetes, standard stats might indicate only one.

On the other hand, making sure to include the truly changing genes comes at the cost of including many false positives. For metabolic diseases, snaring 80 percent of the

true culprits would probably produce a list of more than 13,000 genes — of which more than 12,000 are actually innocent.

People who do these microarray studies have gradually been realizing that standard statistics are not their friend. More sophisticated statistical methods have been developed, and in particular an approach called genome-wide association studies have been used in attempt to deal with some of these issues. It's a step in the right direction, although some problems nevertheless remain.

## Clinical Trials

There is another, much bigger and more serious problem with all these issues about statistics, and that is their use in clinical trials. The double-blind, randomized controlled clinical trial is supposed to be the gold standard for testing new medicines to see whether they are good or bad, and drug approvals and doctor decisions and insurance payments all depend on what these trials find. But clinical trials are fraught with problems. I'm only going to discuss two main points here.

One is the problem of basing conclusions on averages. The other is the myth that that problem is solved by randomization.

When clinical trials are reported, the results are typically given as averages. To give just a crude example, suppose we tested our drug to cure twittering on 1,000 people and compared the results to 1,000 people on placebo. And suppose the drug seemed to cure about 200 people. If about 200 people were also cured in the placebo group, the conclusion would be that there is no statistically significant effect of the drug.

But suppose you look more closely and find that in the drug group, the cure worked on all of the people over age 50. In the placebo group only a few of the over-50s were cured. Because there weren't a lot of over-50 people in the study, they didn't affect the overall average very much. So on average, the drug did not appear to work, so it is not approved, and those poor over-50 people are denied a drug that could alleviate their suffering.

Now you might just say well, let's look at all the subgroups. But that is statistically bogus. By chance, some subgroups will differ. (That's why you sometimes see dumb stories like "Leos are more likely to get in car accidents.") You would have to know in advance which subgroup to look at. Even looking at predetermined subgroups is only a partial solution if it is a solution at all. Because, as I'll discuss in a moment, you will often have no idea what subgroups even exist.

Another example, one of my favorites, involves antidepressants.

A lot of people like to claim that antidepressants really don't work any better than placebos, citing clinical trial results that show there isn't much difference in effectiveness. In fact, most trials seem to show that on average, drugs do work better than placebos, just not by very much. One study published in 2002 analyzed dozens of tests by drug companies seeking that measured depression with a standard questionnaire, with higher scores indicating greater depression. In most trials drugs lowered depression scores more than placebos, but only by an average of about 2

points. The researchers concluded that the difference was small and therefore the clinical significance of antidepressants was dubious.

Of course, the researchers acknowledged that their conclusion hinged on the belief that a drug's true effect is simply the difference between the drug response and placebo response, and that assumes placebos have no effect at all, which is surely wrong. And even if the placebo effect is nil, the notion that drugs aren't worth much remains deeply flawed, because the small reported difference between placebo and drug responses is an average. Every doctor knows that some people may respond to one drug but not to others. Averaging the results may show a small effect only because the large responses for some patients are balanced out by small or no responses from others.

So reporting results as averages is the standard practice, it's what regulators looks at, it's what journalists write about. But it is DUMB. Averages are frequently next to meaningless. A very good piece appeared in *American Scientist* in 2007 that explored this issue in some detail, emphasizing that individual differences are critical in healthcare, not just group differences.


When Averages Hide Individual Differences in Clinical Trials

*Analyzing the results of clinical trials to expose individual patients' risks might help doctors make better treatment decisions*

David Kent and Rodney Hayward

American Scientist. Volume 95
2007 January–February

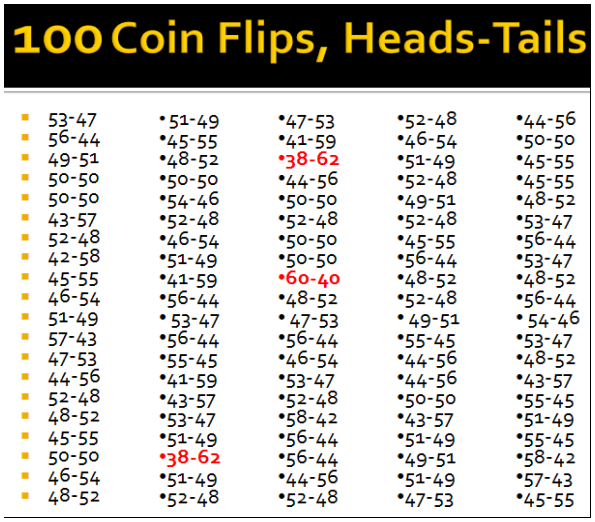
"Reporting a single number gives the misleading impression that the treatment-effect is a property of the drug rather than of the interaction between the drug and the complex risk-benefit profile of a particular group of patients," the authors wrote. "Determining the best treatment for a particular patient is fundamentally different from determining which treatment is best on average."

Now, the standard response to all this is not to worry, the trials are randomized, so individual differences are averaged out. The control group is comparable to the drug group because there are just as many men as women in each, or the same percentages of different races, or ages, and in any case we can control for those things mathematically if we look and see that they are different. And that's true. You can control for things you know about. But you can't control for the things that you don't know about. And you don't know a lot.

So suppose, first of all, that you only had one thing you didn't know about. When you assign people to two groups, you flip a coin, and so that's the same as assuming you'll have a roughly equal number of heads and tails. If you do one big trial, that will probably be the case. But remember the pennies. Most of the time you get close to the same number of heads as tails. But if you do enough trials, sometimes you'll get that one difference distributed not so evenly.

Here's a little example, trials flipping a coin 100 times. Most of the time the results are close, but you can see here and here and here a split of greater 60-40 or greater. In clinical trials, it's the same way. Even if you have to randomize only one

unknown thing, sometimes it would not work. And there are 20,000 clinical trials going on, so you know some of them won't be well randomized. AND THAT'S JUST FOR ONE UNKNOWN THING.




How many unknown things are there, really? Well, something like 3 million — 3 million places where individuals vary by one of the letters in the genetic catalog, the genome. Sure, most of those differences probably don't matter for any given drug or disease. But let's say that just 1 percent of them mattered. That's 30,000 differences to randomize. Or maybe it's only 0.1 percent or 3,000. You won't successfully randomize 3,000 things when you assign people at random to groups in a clinical trial. Some things won't be randomized. And you have no idea what things weren't randomized or which trial was well randomized for those things and which trial wasn't. The whole foundation of the clinical trial system is BOGUS.

Let me be sure to be clear about one thing here. This can work both ways. It means some bad drugs will get approved, it also means some good drugs will not get approved. This is not about any pharmaceutical company conspiracies. It's about how the math works that is commonly used to do scientific inferences about whether drugs work or not.

To be fair, it's not true that nobody realizes this. Some experts would say that of course, randomization doesn't distribute every characteristic equally between the two groups. The idea seems to be, though, that those unbalanced features will work both directions, some enhancing the effect of the drug, for example, and others impairing it, so it will all balance out. Well, even in an ideal world, the best you could say is sometimes that might happen and sometimes it might not, and you can't know which times are which. But it's also important to remember that the world is not ideal. Protocols are not always followed precisely. Several studies suggest that despite alleged "blindness," clinicians conducting trials sometimes manipulate which patients get the drug. These and other factor can create "selection biases" that distort the findings. "Some of the benefits ascribed to randomization, for example that it eliminates all selection bias, can better be described as fantasy than reality," write Vance Berger and Sherri Weinstein in the journal *Controlled Clinical Trials* (2004). "Selection bias can … cause substantial, systematic and reproducible baseline imbalances … and offer a plausible alternative explanation for observed post-treatment between-group differences."

Furthermore, there are many other problems that afflict clinical trials, such as who is getting randomized to begin with. The people who enter clinical trials may not

be representative of all patients, so the results may not be applicable to the population at large. "The randomization process can serve as a basis of inference," write Lu Zheng and Marvin Zelen of the Harvard School of Public Health. "However, the inference requires that subjects entering a trial constitute a random sample of subjects from a well defined population. Unfortunately this is rare. Subjects entering a trial are not a random sample of patients." The way that results from different centers in a trial conducted at multiple locations also can influence in the conclusions, they point out.

## Meta-analysis

Now, the standard response to all this is that you don't rely on one trial alone. You should combine trials to get a bigger sample and get better statistics. This approach is called a meta-analysis. And guess what. It's also bogus.

In principle, meta-analysis is a good idea. The premise is that if an effect is small (but important, as a fatal side effect in some small percentage of patients), standard statistics will not regard it as significant unless your trial size is very large.

Here's an example. Let's suppose that someone believes aspirin is deadly for a small number of people in the population, causing them to die from twittering. A study is done on two groups of 100 people each. One group is given a daily aspirin, the other group gets a placebo. After an appropriate period of time, we count up the deaths from twittering and find:

- Aspirin group, 4 die from twittering.
- Placebo group, 2 die.
- Relative risk of aspirin = 2.0 (aspirin doubles the risk)

BUT a statistical significance test finds that the relative risk at the 95 percent confidence interval ranges from 0.37 – 10.7, so this increase is not statistically significant. And it is a small risk. If 2 percent of the population dies from twittering anyway, then only 2 percent more die if they take aspirin. But because so many people twitter, and so many take aspirin, that 2 percent could add up to a lot of deaths. So this is a case where a small effect could be significant, even if not statistically significant. And in this case, for an increase that small to be statistically significant, you'd need about 600 people in each group.

So if all the evidence you have is from small trials, none of them would be likely to find a statistically significant effect.

By combining several small trials, though, the sample size could become big enough to detect the effect. But is it valid to combine trials together like that? Will the mathematics that make statistical tests possible still apply? Yes, if some criteria are met:

1. You have to find ALL the trials that have been done on this issue.

2. They all need to have been done using the same protocols, same definitions of terms, involving the same kinds of people and same kinds of controls, with outcomes measured in the same way.

The trouble is, these criteria are rarely met

A couple of years ago a meta-analysis of echinacea indicated that it was effective at treating the common cold. A newspaper account of that report included the following comment from an outside expert, Dr. Bruce P. Barrett of the University of Wisconsin. "If you're testing the same intervention on the same population using the same outcome measures, then meta-analysis is a very good technique," Dr. Barrett said. "But here every one of those things fails."

In other words, this meta-analysis was bogus. And these requirements frequently aren't met in other meta-analyses either.

I'll mention two other examples of bogus meta-analysis just to show you how serious this problem can be. One involves Avandia, a drug for treating diabetes. The other involves the risk of suicide for kids given antidepressants.

A paper in 2007 in the *New England Journal of Medicine* claimed that Avandia "significantly increased the risk of heart attacks," one newspaper reported. "Researchers also found that the drug boosted the chances of dying of heart disease by 64%," wrote another. Those writers did not, however, appear to have much of a clue about the evidence underlying their statements – a good example of the confusion between "significance" and "statistical significance."

In fact, the reported increase in deaths was NOT statistically significant (at the 5 percent level). The increased heart attack risk was statistically significant, but just barely.

One newspaper report even went so far as to provide the raw data, noting that 86 heart attacks occurred among 15,560 patients who took Avandia, while only 72 heart attacks were recorded among the 12,283 given a placebo or some other drug. That makes it sound like many more heart attacks occurred with Avandia. But if you pause to do a little elementary math, you would notice that Avandia was taken by more people. A proper comparison would point out that 59 people per 10,000 had heart attacks without Avandia, and only 55 people in 10,000 had heart attacks when using Avandia.

So that makes it look like Avandia is a good thing. But that was before the statistical analysis began. These numbers came from a combination of dozens of clinical trials, most of them small, with only a few hundred patients. To do a meta-analysis of many trials, you don't just add the numbers up. You have to follow the statistical rules for how to combine the studies. It's a technical process, involving several mathematical manipulations for combining the studies and deciding how much weight to give to each of them. At the end of that process, the numbers suggested that Avandia actually increased the risk.

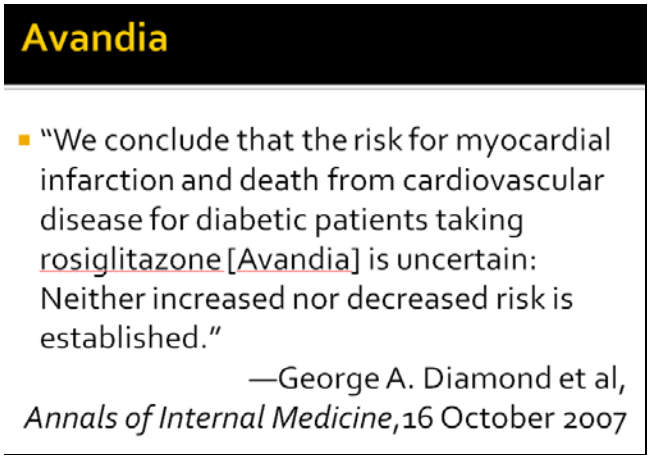
It's certainly possible that even though the raw numbers show a benefit, proper statistical analysis may reveal an actual risk. But for those statistical manipulations to

be valid, the criteria for doing a meta-analysis must be met. All the studies ever done must be included, unpublished as well as published. And all studies must use the same methods, the same definitions and procedures.

Many unpublished studies were in fact included in the Avandia meta-analysis. But they were not all conducted according to the same rules. In some cases, Avandia was given along with other drugs. Sometimes the non-Avandia group got placebo pills, while in other trials that group received another drug. And there were no common definitions.

"Across the trials, there was no standard method for identifying or validating outcomes; events . . . may have been missed or misclassified," Drs. Bruce Psaty and Curt Furberg wrote in an editorial accompanying the *New England Journal* report. "A few events either way might have changed the findings…. In this setting, the possibility that the findings were due to chance cannot be excluded."

Unfortunately, you can't conclude that chance IS to blame, either. It may be that Avandia really does raise the risk of heart attack, and even if the risk is small — a few extra attacks per thousand people or so — the consequences are serious when a million people are taking the drug. But this meta-analysis does not establish the risk very convincingly (as a paper appearing soon after the original meta-analysis made clear).




Very similar issues arose with studies claiming to show that drugs for treating depression called SSRI's cause kids to commit suicide. About five years ago now the Food and Drug Administration concluded from a meta-analysis that it was necessary to require a "black box" warning of the drugs' dangers. Actually, the evidence for that conclusion was probably the shoddiest example of statistical reasoning I have ever seen, and I've seen some shoddy ones. And media coverage of the issue was about as reliable as AIG's TV commercials.

Now it may very well be that in some children, SSRIs do cause a tragic bad reaction that induces suicide, and it is prudent to warn of that possibility. Another danger, though, is that such warnings may become so shrill that seriously sick youth forgo the best treatment available to help them. And untreated depression, there is no doubt at all, often does end in suicide. And in fact, after years of declining, the suicide rate in adolescents did begin to rise again after that black box warnings were imposed.

The problem in this case is that people with depression sometimes contemplate suicide whether taking medication or not. The issue is whether taking medication makes suicidal behavior more likely than it would otherwise be.

As it turns out, in all the studies the FDA analyzed (more than 20 studies of drugs for treating depression in youth ages 6-18), for all the drugs and all the disorders, the

number of suicides reported was precisely zero. The conclusion of danger was based on suicide attempts or plans or reports of suicidal thoughts. For instance, in three clinical trials of Prozac for depression, 17 cases of such suicidal behavior or thinking were recorded (out of 355 patients in the trials). But of those 17 cases, only eight were among patients given Prozac. Among those given placebos, nine such incidents were recorded.

For some other antidepressants, suicidal incidents were reported more often for patients taking the drug than for those on placebo. But in each study the difference was again too small to be statistically significant. But of course, lack of statistical significance does not mean there is no effect; it could be that the effect is just too small to detect. That's why the FDA lumped all the studies together, weighing some more heavily than others based on the number of patients and the number of incidents.

But when the studies of SSRIs for treating serious depression were lumped together, there was still no statistical significance. The only way the report reached a conclusion of statistical significance was to add in studies of NON-SSRIs, and then to add in studies using SSRIs for diagnoses other than depression. Adding all those studies together, in a certain way, it appeared that, just barely, the drugs might cause a slight increase in suicidal thoughts or behavior in youth beyond that expected by chance. But for the SSRIs alone, when used for depression, the excess of incidents with drug over placebo was not statistically significant.

Still, for some of the SSRIs, the number of incidents was a little higher with the drug than the placebo. Those SSRIs show what the FDA calls a "signal" — an excess, possibly meaningful, but small enough to be possibly due to chance. With Paxil, for example, trials recorded more than twice the rate of incidents on the drug than on placebo. For Prozac, the signal is negative, with more suicidal incidents on the placebo than the drug. So it looks like Paxil is worse than Prozac.

But guess what. The rate of suicidal incidents was higher with Prozac. (In the Paxil/depression trials, 3.2 of every hundred patients on the drug reported suicidal incidents; for Prozac, it was 4.5 per 100.) The apparent safety advantage of Prozac over Paxil was due not to the behavior of kids on the drug, but to kids on placebo — fewer kids on placebo in the Paxil trials reported incidents than in the Prozac trials. It just goes to show how slippery statistics can be, and shows how foolish it is to base policy on data so disturbingly dubious.

### All is not wrong

So, the situation is indeed grave. But I don't want to leave the impression that all science is worthless, or that we actually know nothing about anything.

After all, in many scientific studies the results have a very high level of statistical significance. If a result is statistically significant at the 0.00001 level, it's much less likely to be spurious than those results at the 5 percent level. The cautions still apply, but the prospect for serious misinterpretation is lessened. In fields like particle physics, for instance, researchers insist on that sort of confidence — they yawn when shown a result at the 5 percent level and would ridicule anybody who tried to publish it.

Furthermore, when a result from one study is replicated, and replicated again, the statistical confidence in the result grows dramatically. Scientists know that, and they often warn that results have to be replicated to be accepted. And journalists even sometimes mention that in their stories. But they write the stories nonetheless as though the new finding is ready for the textbooks.

Scientists studying gene links to disease have recognized this and that is why the method called genome-wide association that I mentioned earlier has become popular. It builds in an internal replication process to help indentify disease-linked genes more effectively.

Finally, there is one other major cause for hope, and that is the renaissance of another approach to statistics that addresses many of the shortcomings of the standard approach. I'm talking about Bayesian statistics.

## Bayesian statistics

The statistical philosophy I've been talking about so far is called the frequentist approach. Bayesian and frequentist methods have historically been in conflict. Bayesian methods stem from a posthumously published paper in 1763 by the English clergyman Thomas Bayes. In a Bayesian analysis, probability calculations require a prior value for the likelihood of an association, which is then modified after data are collected. When the prior probability isn't known, it must be estimated, leading to criticisms that subjective guesses must often be incorporated into what ought to be an objective scientific analysis.

Let me offer a simple example of Bayesian reasoning at work.

Suppose the test for steroids is 95 percent accurate (that is, it correctly identifies steroid users 95 percent of the time, and identifies non-users as users 5 percent of the time). So it's the analog of the 5 percent P value for drawing a conclusion.

An anonymous player (let's call him Barry) tests positive. What is the probability that he really is using steroids?

If you say 95 percent, you are a frequentist, or at least a frequentist sympathizer. But remember my point that to draw a sound conclusion, you need to know some facts not included in this evidence. In this case, you need to know how many baseball players use steroids to begin with. That would be what a Bayesian would call the prior probability.

Let's say it has been established that 5 percent of major league baseball players use steroids. Now suppose you test 400 players. How many would test positive?

- Out of the 400 players, 20 are users (5%) and 380 are not users.
- Of the 20 users, 19 would be identified correctly (95%).
- Of the 380 non-users, 19 would incorrectly be indicated as users (5%).

So if you tested 400 players, 38 would test positive; 19 would be guilty users and 19 would be innocent non-users.

So when you test a player and his test is positive, the chances that he really is a user are 50 percent.

Now that's a simple trivial example; nobody cares if baseball players are taking steroids. But the same principle applies in serious situations, like medical diagnoses.

Suppose you have just tested positive for a rare but always fatal disease that afflicts 1 person in a million. The test is 99 percent accurate.

Should you:

A. Sell your stock and go skydiving.
B. Confess all your sins on your blog.
C. Request results of the B sample.

If you know Bayesian statistics, you won't worry too much. In this case the prior probability of having the disease is 1 chance in a million. For a 99 percent reliable test, there will be 10,000 false positives after testing a million people and only one correct positive. So the odds that you have the disease are 1 chance in 10,000.

Bayesianism's star has risen and fallen over the past two centuries, and it fell mostly out of favor among 20th century biostatisticians after the ascendancy of the frequentist approach reflected in statistical tests formalized by Fisher and Neyman and Pearson. In recent years, though, Bayesian methods have gained favor among many researchers. Berry of M.D. Anderson has been advocating its revival for decades. Recently Bayesian methods have become more popular in clinical trials, and have been frequently advocated for improving the conclusions reached in gene association studies.

Now, it is true that Bayesian statistics has its problems, too. There is no magic cure for all the ills of math in science and medicine today. But the situation would be a lot better if more scientists, and journalists, understood these ills to begin with and took steps to cope with them more intelligently. In particular, journalists need to learn to ask questions like how many other labs might be working on a question, what does the previous evidence indicate, is there a reason why a newer study is any more reliable than an older study. And reporters need to learn how to phrase the meaning of statistical results so as not to be trapped in logical or mathematical errors.

Finally, I think scientists and journalists alike need to listen to the words of Fisher himself. All I am saying here today is that sound judgments require … sound judgment, and not the blind devotion to numbers alone that has damaged so much of modern life in so many different ways.

Fisher said that the wise scientist draws conclusions not by numbers alone, but "rather gives his mind to each particular case in the light of the evidence and his ideas." And I think that's a good idea.

I'll end with a couple of observations that others have made that illustrate, I think, why all this technical stuff is worth paying attention to, and in particular why journalists should be interested in it. One is from a very nice popular book on statistics,

which midway through, almost as an aside, offered a comment that scientists ought to take note of:

"What does probability mean in real life? . . . This problem is still unsolved, and ... if it remains unsolved, the whole of the statistical approach to science may come crashing down from the weight of its own inconsistencies" (David Salsburg, *The Lady Tasting Tea*).

The other is from a statistician whose papers have been making a lot of these points, particularly the point that others need to pay attention.

"In a world where medical researchers have access to increasingly sophisticated statistical software, the statistical complexity of published research is increasing, and more clinical care is being driven by the empirical evidence base, a deeper understanding of statistics has become too important to leave only to statisticians" (Goodman 1999).

## References and Additional Reading

### Articles

Berger, Vance W. and Sherri Weinstein, "Ensuring the comparability of comparison groups: is randomization enough?" *Controlled Clinical Trials* 25 (2004).

Berry, Donald A. "Bayesian Clinical Trials." *Nature Reviews Drug Discovery* 5 (January 2006), 27-36. A thorough presentation of the case for using Bayesian statistics in medical research.

Goodman, Steven N. "Toward Evidence-Based Medical Statistics. 1: The *P* Value Fallacy." *Annals of Internal Medicine* 130 (15 June 1999), 995-1004. A detailed and clear discussion of the reasons why P values are commonly misinterpreted.

Hubbard, Raymond and J. Scott Armstrong, "Why We Don't Really Know What 'Statistical Significance' Means: A Major Educational Failure," *Journal of Marketing Education* 28 (August 2006).

Ioannidis, John P. A. "Why Most Published Research Findings Are False." *PLoS Medicine*, 2 (August 2005), 0101-0106. An assessment of the many factors casting doubt on the conclusions drawn in most medical studies.

Kent, David and Rodney Hayward, "When Averages Hide Individual Differences in Clinical Trials." *American Scientist* 95 (January-February 2007).

Stephens, Philip A. Steven W. Buskirk and Carlos Martínez del Rio, "Inference in ecology and evolution," *Trends in Ecology and Evolution* 22 (2007).

Stroup, T. Scott et al. "Clinical Trials for Antipsychotic Drugs: Design Conventions, Dilemmas and Innovations." *Nature Reviews Drug Discovery* 5 (February 2006), 133-146. An in-depth discussion of the statistical and methodological issues afflicting clinical trials, focusing on antipsychotic drug testing.

Sullivan, Patrick F., "Spurious Genetic Associations," *Biological Psychiatry* 2007, doi:10.1016/j.biopsych.2006.11.010

Wacholder, Sholom, Stephen Chanock, Montserrat Garcia-Closas, Laure El ghormli, Nathaniel Rothman, "Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies," *Journal of the National Cancer Institute* 96 (March 17, 2004).

Zheng, Lu and Marvin Zelen, "Multi-center clinical trials: Randomization and ancillary statistics," *Annals of Applied Statistics* 2 (2008).

**Books**

Jenkins, Stephen H. *How Science Works: Evaluating Evidence in Biology and Medicine*, New York: Oxford University Press, 2004. A series of case studies illustrating the uses and misuses of math in testing hypotheses and quantifying results in medical and environmental research.

Salsburg, David. *The Lady Tasting Tea*. New York: W.H. Freeman, 2001. An engaging account of 20th-century developments in statistical science.

Weaver, Warren. *Lady Luck: The Theory of Probability*. Garden City, N.Y.: Doubleday, 1963. An old but entertaining introduction to the mathematics of probability and statistics.