# Warning Signs in Experimental Design and Interpretation

When an experimental study states "The group with treatment X had significantly less disease ($p$ = 1%)", many people interpret this statement as being equivalent to "there is a 99% chance that if I do treatment X it will prevent disease." This essay explains why these statements are not equivalent. For such an experiment, all of the following are possible:

- X is in fact an effective treatment as claimed.
- X is only effective for some people, but not for me, because I am different in a way that the experiment failed to distinguish.
- X is ineffective, and only looked effective due to random chance.
- X is ineffective because of a systematic flaw in the experiment.
- X is ineffective and the experimenters and/or reader misinterpreted the results to say that it is.

There is no way to know for sure which possibility holds, but there are **warning signs** that can dilute the credibility of an experiment. In Part I we look at warning signs in the design of an experiment that can render it uninformative; in Part II at warning signs in the interpretation of an experiment that can lead the reader to give it more credibility than it deserves. The presence of any one warning sign does not invalidate a study -- there certainly are valid and convincing studies that are not randomized, for example -- but the more warning signs the more skeptical you should be.

# Part I: Common Warning Signs in Experimental Design

## Warning Sign D1: Lack of a Randomized Controlled Trial

The most reliable experiment to evaluate a medical treatment is a randomized controlled trial, in which a population is randomly divided into a *test group*, which receives the treatment, and a *control group,* which does not.

Why are controls important? Because you want a fair comparison. If you choose subjects who are unusually healthy and give them a treatment, you may well get unusually positive results, but it won't tell you much about the treatment. You have to think of all the variables that must be controlled for: temperature, pressure, age of subjects, prior history of disease, etc, and make sure that your subjects and your controls are balanced on these variables. You might think that in some cases you can use historical values: in the general population, x% are prone to get disease D, and they recover in y days on average. Using such historical figures is fraught with danger, because you don't know how your population of subjects or your experimental conditions differ from the historical norms. It is far safer to use a real control group, as well as a test group. As Geneticist Gerald Fink (quoted by Natalie Angier) says "*In my life as a scientist, the thing I worry about the most is, What are the right controls? You send a paper off for publication, and you're stricken with doubt: Did I do it? Did I use the right controls?*"

How do you achieve a good balance between the test and control group? One good way is to *randomize* the subjects. You can do this by randomly assigning all the subjects to one group or the other, or by first stratifying them into similar subsets, and then randomly assigning members from each subset into the test and control groups. It wouldn't do for the experimenter to assign subjects to one group or the other after having met the subjects; the experimenter might (consciously or unconsciously) assign healthier patients to one group.

In some cases, you can assign all subjects to both groups: in psychology experiments that measure things like reaction time, you can every subject try every experimental condition (in randomized order), because there is little or no influence between each trial. Obviously, you can't have subjects try every possibility on things like cardiac operations: you can only choose one option for each subject.

Of course there are situations where you can't use a randomized controlled trial. You can't ethically test the effects of cigarettes by requiring a control group to smoke. In such situations, you make do with non-randomized studies, and there are various statistical techniques for dealing with the situation. But if you see a published study that *could* have used a randomized controlled trial and didn't, that's a warning sign that something may be wrong.

## Warning Sign D2: Lack of Double-Blind Studies

A *blind* study is one where the subjects don't know what group they are in. A *double-blind* study is one where the experimenters don't know either. Why is this important? We know there is a *placebo effect* wherein patients do better when they are told they are receiving a treatment: the patients' expectations play a role in their recovery. To make sure we are studying the effect of the treatment itself and not the patients' expectations, it is better to give all patients the same expectation. So we tell them, for example, "take this pill, it might be experimental drug X or it might be a sugar pill." The double-blind part is important because we don't want the experimenters to subconsciously tip off the subjects as to what group they are in, nor to treat one group differently than the other, nor to analyze the results differently.

## Warning Sign D3: Too Few Subjects

A well-structured randomized controlled trial eliminates systematic biases, but is still open to random variations. It is always possible that by random chance, all the really healthy subjects get assigned to the test group and all the sick subjects get assigned to the control group, thereby making a worthless treatment look good. Since we can't eliminate this possibility, statisticians develop a way to measure its likelihood. When we say "the treatment is effective with statistical significance $p$=1%" it means that if there were no difference between the treatment group and the control group, we would expect to see results like this (favoring the treatment group) 1% of the time, just because of the randomized assignment of subjects to one group or the other. The chances go up if there are too few subjects, relative to the measurements being taken.

Psychologist Seth Roberts, who is well-known for experiments with N=1 subject (himself) points out that it is also possible to make the mistake of having too many subjects -- wasting time trying to get confirmation with a large experiment when you should be doing more exploration with smaller experiments to get better ideas that can then be confirmed later. He is quite right that this is a potential problem, and he may well be right that more scientists, in their daily scientific lives, make the mistake of too many subjects than too few subjects. Many of the other points in this essay are also like that: perhaps scientists are *too* cautious about making the leap from correlation to causation. However, when it comes time for a reader to evaluate a published study, we are doing a different task than a scientist trying to come up with an idea, and it is a warning sign when there are too few subjects.

## Warning Sign D4: The Wrong Subjects


Alf Landon

In 1936 *Literary Digest* magazine polled 10 million people and predicted that Alf Landon would win the election for president with 57% of the popular vote. With 10 million responses they certainly did not make the "too few subjects" mistake, but nobody remembers president Landon because he in fact lost 46 of 48 states to Franklin Delano Roosevelt, the 32nd president. Where did *Literary Digest* go wrong? They got their subjects from three sources: their own readers, automobile registrations, and telephone subscribers. But in 1936, as the country emerged from the great depression, many people could not afford literary magazines, nor cars, nor telephones, and those who could not were more likely to vote for Roosevelt. Even today, when phone ownership is much closer to universal, it is difficult to get a truly representative sample of voters. Call a home number during the day, and you get mostly people who don't work. Since there are legal restrictions on calling cell phone numbers for polls, you will underweight the younger voters who tend to prefer cell phones. Pollsters have ways for balancing out these biases, but the sample will never be completely unbiased.

As another example, many psychology experiments are run on volunteer college students. Often an experimenter gets a result from such an experiment and claims it is valid across all people, but later finds out that the result only holds for (a) people roughly 20 years old or (b) people with the skills and dedication necessary to be a college student or (c) the type of people who like to volunteer for things.

## Warning Sign D5: The Wrong Questions

Even a careful study can end up measuring the wrong thing. You do a study that you believe shows that treatment X is effective in relieving stress, and in fact it turns out that it is only effective against the artificial stressful situation you set up in the lab, but not for real stress in the real world. Or, you treat different plots of crops with different fertilizers and believe that the fertilizer used in plot #4 is most effective, when actually wind and water seepage have pushed all the fertilizer south by one plot, and really the fertilizer used in plot #3 is the effective one.

Sometimes experimenters deceptively ask the question they want answered, not the most useful question. In 1987, John Cannell surveyed the results of school testing and found that 50 out of 50 US states reported their children were above average. He called this the "Lake Woebegone Effect." How was it achieved? It turns out that when a school district or state contracts to have its students take a standardized test, they also purchase rights to the scores of a "comparison group," and it appears that test vendors are competing in part on just how low a comparison group they can offer.

## Warning Sign D6: The Wrong Statistics

As an experimenter, the first thing you should do after collecting your data is to *look* at it. Plot the data. Calculate means and standard deviations. Are there any outliers? If so, are they valid measurements, or indications of some error in data collection or recording? Is the data normally distributed? If not, make sure you don't use a statistical test that relies heavily on an assumption of normality. Is the data linear? If not, don't run a linear regression. Are the data points independent? For example, did you take 100 measurements from 100 different subjects, or 10 measurements, one per day, for 10 different subjects? If they are not independent, make sure you use a test that recognizes that.

## Warning Sign D7: Lack of a Specific Hypothesis, or Overzealous Data Mining

A recent article in London's *Sunday Times* reported that clusters of cancer centered around seven different mobile telephone masts, raising the concern that mobile phone radiation might cause cancer. However, there are 47,000 mobile phone masts in

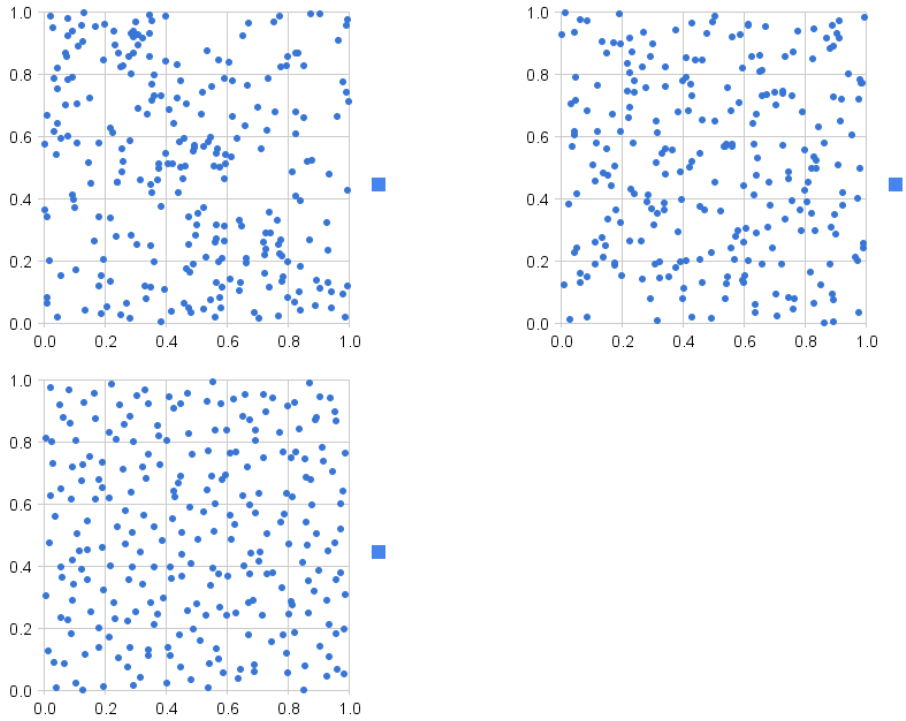
England. The article was titled *Cancer clusters at phone masts* but for all the details it contains it could just as well have been titled *Cell phone masts prevent cancer clusters 99.985% of the time*. It is certainly legitimate to be concerned about these seven cancer clusters, but before concluding that cell phones have anything to do with it, we'd need more data: how many of the other 47,000 masts have cancer clusters? How does that number compare to what we would expect by chance?

Humans are very good at detecting patterns, but rather poor at detecting randomness. We expect random incidents of cancer to be spread homogeneously, when in fact true randomness results in random clusters, not homogeneity. We need careful statistical tests to distinguish between random and non-random clusters. Similarly, if we collect 40 different variables for each subject, it is a mistake to ignore the measurements that show no significant results, and report only on the ones that do. Out of 40 different variables you would expect 2 to have a significant effect at $p$=5%. A report showing 2 such variables is useless: it is indistinguishable from random chance.

A proper experiment states its hypothesis *before* gathering evidence and then puts the hypothesis to the test. Remember when you did your seventh grade science fair experiment: you made up a hypothesis first ("Hamsters will get fatter from eating Lucky Charms than they will from Wheaties") and then did the experiment to confirm or refute the hypothesis. You can't just make up a hypothesis after the fact to fit the data. Compare what happens when a football team scores, but the score is disallowed because a player was offsides. We don't count the score half-way; we count it as zero, because it breaks the rules. The team's fans are free to use the play as evidence of their team's tremendous offensive potential, and point out that it is inevitable that they will score again. On the other hand, the fans of the other team are free to say it was a fluke, and only happened because the player got an unfair advantage from being offsides. The difference between these two opinions will be settled by playing the remainder of the game and seeing who actually does score. In the same way, a statistically significant result observed in a variable that was not part of the original hypothesis for an experiment counts for zero -- using it would be breaking the rules of science. Supporters get to argue that it is inevitable that the variable is in fact significant, but they have to play the game -- in this case, run another experiment -- to prove it.

Is it really fair to rule out serendipity in experiments? The statistician Stephen Senn once complained "The definition of a medical statistician is one who will not accept that Columbus discovered America because he said he was looking for India in the trial plan." He has a point, and I think you should be free to use your own judgment: if you have good reason to believe a result that was not part of the original protocol for an experiment, go ahead and believe it. But don't consider it proved beyond a reasonable doubt and don't insist that others believe it until you have independent studies confirming it.

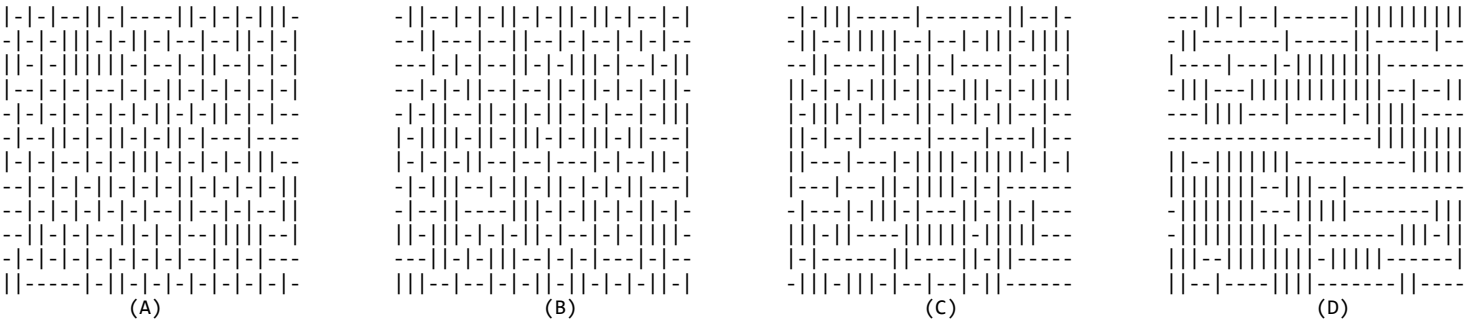
It is interesting to consider *why* people are so prone to see patterns in the data, like the cancer clusters around cell phone masts. It turns out that people and other mammals are sensitive to patterns, and are quick to spot them where they exist, and even when they don't exist. On the other hand, people are poor at identifying randomness. Consider the following three plots. In one of the plots each of the blue points is sampled with equal probability from the entire square. Which one is it?




Most people say the rightmost plot is "most random". Those with some statistical sophistication suspect it may be *too* random, and pick the middle plot. In fact, it is the leftmost plot that is a uniform random sample of 250 points on the unit square. In the middle plot, the grid is divided into the 25 squares shown by the light lines, and ten points are placed (with a random uniform distribution) in each of the 25 squares. The plot on the right is formed in a similar way, except it is composed of 64 smaller squares (not shown by lines), each of which has 4 points placed at random. People don't like the leftmost plot because it has

several clumps of points that seem non-random. In fact, true randomness consists of a mixture of clumps and non-clumps. Randomness is different from homogeneity.

Here's another example: guess which of these four squares was generated with each tick mark independent of the one before it in left-to-right top-to-bottom order?




I asked 10 people; 2 said A, 6 said B, 2 said C, and none said D. Everyone agreed that D has too many long runs of horizontal or vertical marks. Of the others there was disagreement, but the majority said B was about right: not too many long runs, nor too few. In fact C is the correct answer; in C each tick mark is the same as the one before it with probability 1/2. In A the marks flip from one direction to the other 3/4 of the time; in B 2/3 of the time, and in D 1/4 of the time. Presumably my subjects didn't like C because of the several long runs of 6 or 7 ticks in a row. But a run of 7 tick marks should occur with probability 1/64, so in a sample of 300 tick marks, you should expect about 4 or 5 of them. Yet when they do occur, people see it as a pattern, not as randomness.

## Warning Sign D8: Lack of a Theory

The UK's National Radiological Protection Board stated that it considers mobile phones safe in relation to cancer. "Radio waves do not have sufficient energy to damage genetic material in cells directly and therefore cannot cause cancer." They have a *theory* about how radio waves might cause cancer. (It was this theory (not Relativity) that won the Nobel Prize for Einstein, and so it is said that Einstein proved cell phones do not cause cancer.) As long as this theory is correct, statistical correlations about cancer clusters and cell-phone masts do not matter. (Of course, statistical results might cause the Board or others to re-evaluate the theory.) Having a theory is part of having a specific hypothesis, and in fact it is *impossible* to do a proper experiment without at least a skeleton of a theory. Why not? Can't an experiment test the hypothesis that treatment X is effective against disease D, but I don't know why? Actually it is fine to propose that "X prevents disease D in a way similar to other treatments". That counts as a partial theory. Suppose you did a trial of treatment X and found no effect. But a critic says "that's because you did the experiment on a Thursday, and X doesn't work on Thursdays". Or "X only works for subjects wearing purple shoes who have an even number of vowels in their middle name." The critic has a different theory of how the treatment works, and thus of what variables must be controlled for. Without some such theory, you can't do *any* experiment.

## Warning Sign D9: Lack of controls

OK, we know a randomized controlled trial is best, but exactly what should we control? That's a difficult question, answered only by experience. Consider this passage, from Richard Feynman's 1974 commencement address on Cargo Cult Science:

> There have been many experiments running rats through all kinds of mazes, and so on--with little clear result. But in 1937 a man named Young did a very interesting one. He had a long corridor with doors all along one side where the rats came in, and doors along the other side where the food was. He wanted to see if he could train the rats to go in at the third door down from wherever he started them off. No. The rats went immediately to the door where the food had been the time before. The question was, how did the rats know, because the corridor was so beautifully built and so uniform, that this was the same door as before? Obviously there was something about the door that was different from the other doors. So he painted the doors very carefully, arranging the textures on the faces of the doors exactly the same. Still the rats could tell. Then he thought maybe the rats were smelling the food, so he used chemicals to change the smell after each run. Still the rats could tell. Then he realized the rats might be able to tell by seeing the lights and the arrangement in the laboratory like any commonsense person. So he covered the corridor, and still the rats could tell. He finally found that they could tell by the way the floor sounded when they ran over it. And he could only fix that by putting his corridor in sand. So he covered one after another of all possible clues and finally was able to fool the rats so that they had to learn to go in the third door. If he relaxed any of his conditions, the rats could tell. Now, from a scientific standpoint, that is an A-number-one experiment. That is the experiment that makes rat-running experiments sensible, because it uncovers the clues that the rat is really using--not what you think it's using. And that is the experiment that tells exactly what conditions you have to use in order to be careful and control everything in an experiment with rat-running. I looked up the subsequent history of this research. The next experiment, and the one after that, never referred to Mr. Young. They never used any of his criteria of putting the

> corridor on sand, or being very careful. They just went right on running the rats in the same old way, and paid no attention to the great discoveries of Mr. Young, and his papers are not referred to, because he didn't discover anything about the rats. In fact, he discovered all the things you have to do to discover something about rats. But not paying attention to experiments like that is a characteristic example of cargo cult science.

Feynman is saying that Young figured out all the things you need to control to have a good rat-maze experiment. Unfortunately, many researchers have failed to adhere to all these controls.

I'm reminded of when I took an undergrad psychology course, and one of the requitrements was to be a subject in an experiment. I forget the exact task, but the setup was that some stimulus was flashed on a screen, and then I was supposed to press one of four buttons corresponding to the right answer as fast as possible. In these days the apparatus was not fully computer-controlled, so one experimenter announced the trial number beforehand, then the stimuls was presented and I responded, and another experimenter would record the result manually on a clipboard. But after hearing a few of the properly-randomized trial numbers: 17, 32, 3, 26, ... it occurred to me that the correct button number was always the trial number modulo 4. So from then on I didn't even have to look at the stimulus, I could just listen for the trial number, say 43, quickly compute that that 43 = 3 mod 4 and then press button 3 as soon as the stimulus pops up. The experimenters thought they had correctly randomized the order of presentation of the trials, but they had neglected the fact that announcing the trial number beforehand spoils the whole thing (it would have been okay for the first experimenter to announce the trial number after I had pressed the button).

# Part II:
# Common Warning Signs in Interpretation of Experiments

## Warning Sign I1: Lacking Repeatability and Reproducibility

If an experiment indicates a phenomenon that is in fact real, then the experimenter should be able to *repeat* the experiment and get similar results. More importantly, other researchers should be able to *reproduce* the experiment and get similar results as well. Gullible people get excited by the very first result, but wiser heads wait for reproducible evidence, and don't get fooled as often by false alarms.

## Warning Sign I2: Ignoring Publication Bias

Here is my amazing claim: under the strictest of controls, I have been able, using my sheer force of will, to psychically influence an electronic coin flip (implemented by a random number generator) to come up heads 25 times in a row. The odds against getting 25 heads in a row are 33 million to 1. You might have any number of objections: Is the random number generator partial to heads? *No.* Is it partial to long runs? *No.* Am I lying? *No.* Do I really have telekinetic powers? *No.* Is there a trick? *Yes.* The trick is that I repeated the experiment 100 million times, and only told you about my best result. There were about 50 million times when I got zero heads in a row. At times I did seem lucky/telekinetic: it only took me 2.3 million tries to get 24 heads in a row, when the odds say it should take 16 million on average. But in the end, I seemed unlucky: my best result was only 25 in a row, not the expected 26.

Many experiments that claim to beat the odds do it using a version of my trick. And while my purpose was to intentionally deceive, others do it without any malicious intent. It happens at many levels: experimenters don't complete an experiment if it seems to be going badly, or they fail to write it up for publication (the so-called "file drawer" effect, which has been investigated by many, including my former colleague Jeff Scargle in a very nice paper), or the journal rejects the paper. The whole system has a publication bias for positive results over negative results. So when a published paper proclaims "statistically, this could only happen by chance one in twenty times", it is quite possible that similar experiments *have* been performed twenty times without a positive result, but have not been published.

## Warning Sign I3: Ignoring Other Sources of Bias

In any statistical sample there are two possible sources of error: variance and bias. The variance is the random fluctuation due to the fact that we can sample only a small part of the total population. We measure the probability of a variance error with the *p* score (more on that later in Warning Signs I5 and I6). There are many possible sources of bias errors (for example, see Warning Sign D4), but they cannot be neatly quantified with a numeric score, so there is a tendency to ignore them. But bias is still there, and ignoring it means that more results are accepted that should not be. In fact, John P. A. Ioannidis went so far as to claim that Most Published Research Findings are False. This editorial is a bit misleading in its title, because it does not actually count how many studies are wrong. In fact, it makes no claims whatsoever about the efficacy or shoddiness of experiments. Rather, what it does is demonstrate a *mathematical* claim that under certain assumptions about the degree of bias and the percentage of true

correlations compared to total possible correlations in a field of study, then for a given *p* value you can estimate how many published results are in fact true. He gives an example of looking at a set of 100,000 gene polymorphisms to see which are asssociated with schizophrenia. Under a set of reasonable assumptions, if the bias is 10%, then the probability of any association that is reported to be significant at the *p*=5% level is actually only 0.044%. This is related to our Warning Sign D7, overzealous data mining: with 10,000 possibilities to choose from, but most of them having no effect, it is more likely to get a result due to random chance than due to true effect.

## Warning Sign I4: Confusing P(H | E) with P(E | H)

In an article titled *An Intuitive Explanation of Bayesian Reasoning*, Eliezer Yudowsky considers this question:

> *1% of women at age forty who participate in routine screening have breast cancer. 80% of women with breast cancer will get positive mammograms. 9.6% of women without breast cancer will also get positive mammograms. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?*

Only about 15% of doctors get this right (See Casscells, Schoenberger, and Graboys 1978; Eddy 1982; Gigerenzer and Hoffrage 1995; and others; researchers keep repeating the study and doctors keep getting it wrong.) Most doctors estimate the probability to be between 70% and 80%. The correct answer is 7.8%. Here's why: consider 1000 women of age forty. 1% of them (10 women) will have breast cancer. Of those, 80% (8 women) will get a positive mammogram. What about the other 990 women? 9.6% of them, or 95 women, will also get a positive mammogram. So there are 95 + 8 = 103 anxious women with a positive result, but only 8 actually have cancer. 8/103 = 7.8%. See the table below:

| | cancer | no cancer | total | P(cancer \| positive) |
|---|---|---|---|---|
| positive mammogram | 8 | 95 | 103 | 8/103 = 7.8% |
| negative mammogram | 2 | 895 | 897 | |
| total | 10 | 990 | 1000 | |

It says right in the problem description that P(positive | cancer) = 80%. (Note: Read "P(positive | cancer)" as "the probability of a positive test result given that the patient has cancer.") But what we're interested in is P(cancer | positive), the probability of cancer given a positive result. This is not the same thing at all, as the computation above shows, but doctors (and lay people) seem to reason that they are more or less the same. In general, statisticians talk about P(H | E), the probability of a *hypothesis* given the *evidence*. In this case, the hypothesis is "having cancer" and the evidence is "positive mammogram." But usually the probability numbers that are available to us are given in terms of P(E | H); in this case P(positive | cancer) = 80%.

To make the distinction more clear, let's consider a much rarer disease, central nervous system vasculitis (CNSV), which affects an estimated 1 in a million people. Let us assume there is a test that is 99% accurate (for both patients with and without the disease). What is the chance that a patient with a positive test result for CNSV actually has it? This time, we'll consider a hundred million patients:

| | CNSV | no CNSV | total | P(CNSV \| positive) |
|---|---|---|---|---|
| positive test | 99 | 999,999 | 1,000,098 | 99/1,000,098=0.01% |
| negative test | 1 | 98,999,901 | 98,999,902 | |
| total | 100 | 99,999,900 | 100,000,000 | |

Even though the test is 99% accurate (that is, P(positive | CNSV) = 99%), a patient with a positive test result has only an 0.01% chance of having the disease (that is, P(CNSV | positive) = 0.01%). This is a very small chance, but it is 100 times higher than patients without a test result, and 10,000 times higher than patients with a negative result.

Now let's go to the other extreme. Assume the probability of having the common cold during cold season is 10%. Again, let's assume a test that is 99% accurate.

| | cold | no cold | total | P(cold \| positive) |
|---|---|---|---|---|
| positive test | 99 | 9 | 108 | 99/108=92% |
| negative test | 1 | 891 | 892 | |
| total | 100 | 900 | 1000 | |

This time the probability of having a cold given a positive test is 92%.

Now consider carefully these two statements:

*"Sam has a positive test result for CNSV, and the test is correct 99% of the time."*
*"Pat has a positive test result for a cold, and the test is correct 99% of the time."*

You will probably feel an overwhelming urge to say that both Sam and Pat have a 99% chance of their respective diseases (or something close to 99%). *You must resist giving in to this urge*, because you now know that Sam's chance is really 0.01% and Pat's 92%. Perhaps the following example will help: suppose I have a test machine that determines whether the subject is a flying leprechaun from Mars. I'm told the test is 99% accurate. I put person after person into the machine and the result is always negative (correctly). Finally one day, I put someone (say, Tom Hanks) into the machine and the light comes on that says "Flying Leprechaun!" Would you believe the machine? Of course not: that would be ridiculous, so we conclude that we just happened to hit the 1% where the test errs. We find it easy to completely reject a test result when it predicts something impossible (even if the test is very accurate); now we have to train ourselves to *almost* completely reject a test result when it predicts something *almost* completely impossible (even if the test is very accurate).


Carl Sagan

Carl Sagan (1934-1996) famously said "*Extraordinary claims require extraordinary evidence.*" (although he was repeating the advice of Marcello Truzzi (1935-2003) and Pierre Laplace (1749-1827), who said "The weight of evidence for an extraordinary claim must be proportioned to its strangeness."). Intuitively this seems like good philosophy, but we have just seen that it is actually a precise statement about *mathematics*, not just a vague claim about philosophy. For an ordinary claim ("Pat has a cold"), evidence in the form of a test that has a 99% chance of being correct is good enough to give you 92% confidence in the result. But for an extraordinary claim ("Sam has CNSV"), a test with 99% probability really doesn't help much; we would need a test with something like 99.999% accuracy before we will start to believe the claim. Prof. Michael Wigler has a more pessimistic way of putting it (quoted by Natalie Angier): "Most of the time, when you get an amazing, counterintuitive result, it means you screwed up the experiment."

## Warning Sign I5: Taking *p* too Seriously

When an experiment states that the results were significant at the $p$=1% level, it means P(Evidence | (Hypothesis is false)) = 1%. (Or, equivalently, that P(Evidence | (Null Hypothesis is true)) = 1%. Either the experiments's hypothesis must be true, or the null hypothesis must be true.) Note that this by itself says nothing directly about P(Hypothesis | Evidence), which is what we really want to know. Note also that the $p$ score is only talking about the chance of an error due to random chance, and says nothing about any of the other sources of mistakes. And yet, time after time, doctors and even trained statisticians see "$p$=1%" and think "There's a 99% chance this result is true." That is a mistake.

In my own work, we do a dozen or more experiments every day, and because we have large data sets our typical value is $p$=0.0000001% -- one in a billion or so. Sometimes we get a $p$ of one in a trillion. And even then we agonize over whether we should believe the results of the experiment. We're not worried about an error due to random chance, but we are worried about other problems: perhaps this experiment works only because it is novel, and the effect will wear off as people habituate to it; perhaps it works in the countries we tested, but won't work in the rest of the world, and so on. Consider the $p$ value as one source of error, but don't ignore the other sources of error.

Don't confuse the *statistical significance* of an experiment with the *magnitude* of the result, even though the word "significance" is often used for both. In my work we often run an experiment that produces a statistically significant improvement, but we don't bother to implement the change because the magnitude of the improvement is small. For example, the experiment might show that alternative X is better than the status quo with statistical significance $p$=0.0000001%, but that the magnitude of the difference is small, perhaps 0.01%. In other words, we are almost certain that X would be better, but it would be better to such a small degree that nobody would really notice any difference. In such a case it might not be worth the expense and complication of switching to the slightly better method.

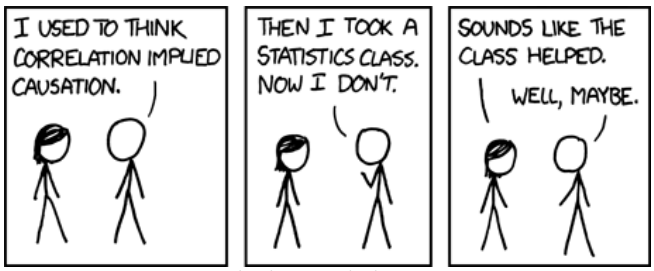
Tradition dictates that you should report your $p$ score, but it is almost always more informative to report a **confidence interval**, to show the magnitude of any effect. For example, rather than just saying that method X is better than method Y with a statistically significant result at the $p$=5% level, you should also say that the 95% confidence interval for X is a score of 327 to 329, while the 95% confidence interval for Y is 329 to 330.

## Warning Sign I6: Accepting the Wrong *p* Value

What $p$ value is sufficient to accept the results of an experiment? It should be clear by now the answer is "it depends." It depends on the prior probability of the Hypothesis. The $p$ value necessary to convince me that Sam has CNSV is much stricter than the one necessary to convince me that Pat has a cold. In High School science classes, $p$=5% is considered good enough, because most of the time we really know the answer anyway, and the experiment is done just to give the students practice. In most physics journals, the standard is $p$=0.01%. The idea is that there are real phenomena to be discovered, and it is better to insist on more careful experiments so that we can be sure. In medical journals, $p$=5% is often accepted, although some insist on $p$=1%. Why are medical journals more lenient? In part because it is harder to get a good result -- animal subjects are more fickle than beakers full of chemicals or wires full of electrons. In part because it is harder to get a good theory. And in part because we want to encourage the dissemination of potentially life-saving information, and we expect doctors to be critical well-informed consumers of this information (an expectation that lamentably appears not to be warranted).

What's the right *p* value? There can't be a single answer. The answer actually does not depend on just statistics and probabilities; it depends on *utility*. That is, we have to answer what expected value (or cost, depending on how you want to look at it) would we get in each of the four possibile cases: the result is true (or false) and we act on it as if we believe it (or don't believe it). For example, treatment X might or might not cure cancer, and we might or might not believe the study that says it does. If there are no potential bad side effects of X, and if using X does not preclude using other potential cures, then we would be inclined to believe that X is effective even with a relatively poor *P* value. If X has powerful negative side-effects, we would insist on more compelling evidence before using it.

## Warning Sign I7: Confusing Correlation with Causation



xkcd: Correlation

Statistical studies can easily show that one variable is *correlated* with another. Proving that one *causes* the other is more difficult. Consider this story about a study of cell phones and brain tumor risk. First, the study is a good example of hysteria: it caused a public outcry about the dangers of cell phone usage. However, if you read the article carefully, you'll see that the main finding does not actually address causation of cancer by cell phones, only correlation. The study looked at people who already had brain tumors. They found that tumors are more likely to occur on the side of the head on which the phone was most often held. But there are three possibilities that would lead to this correlation: (1) Holding the phone on one side causes a tumor, (2) Developing a tumor causes one to hold the phone on that side, and (3) another variable or set of variables causes both. Because it was a long-term study, we can largely rule out (2): the study would have covered the time before tumors developed. But the experiment doesn't distinguish between (1) and (3). For example, assume that 90% of users hold the phone in their right hand. Now we'll make an unwarranted assumption for this exercise: we'll assume that 80% of tumors develop in the right side of the brain and that tumor location and cell phone usage are completely independent of each other. With a sample of 1000 people we'd get this:

| | tumor on R | tumor on L |
|---|---|---|
| hold on R | 720 | 180 |
| hold on L | 80 | 20 |

How many get the tumor on the same side as the phone? (720+20)/1000 = 74%. Only 26% get it on the opposite side, so if you're not careful you might claim that "cell phone usage triples the incidence of brain tumors" when actually (given the assumptions of this exercise) tumor location is completely independent of cell phone usage.

Now let's change the assumptions. We'll stick with 90% right-hand use, but we'll assume that 10% of tumors are caused by the cell phone and appear on the same side, and the other 90% are split evenly between the two sides. With 1000 people we get these expected results:

| | tumor on R | tumor on L |
|---|---|---|
| hold on R | 495 | 405 |
| hold on L | 45 | 55 |

So this time we get (495+55)/1000 = 55% of tumors on the same side, 20% less than last time, but this time there *is* a causation. So if we can't rely on the numbers to distinguish causation from correlation, what can we do? We need a *counterfactual intervention*: to prove *A* causes *B*, we need to have cases where *A* may or may not occur, and then observe that when we intervene to make *A* occur, *B* happens, and when we make *A* not occur, *B* does not happen. The problem with this study is that it looked at people who were known to have cancer: *B* has already happened and we have no way to intervene. One good way to intervene is with a randomized controlled trial. Of course it is difficult to do a randomized trial on topics like this. First, brain tumors are very rare, so you'd need a lot of subjects. Second, it would be difficult to get subjects to go along with the trial: "OK, Ann, you're not allowed to use a cell phone at all for the next twenty years. Bob, you can use one in your right hand, and Charlie, you have to hold it in your left hand. ...".

How do you move from correlation to causation, if you can't do a randomized controlled trial? For example, how can you show that smoking causes lung cancer? I won't go into details here, but you could look at the literature on propensity scores, double robustness, and selection bias. Ideally you'd like to have as many of the following as possible:

1. Observational studies (for example, studies of patients with and without cancer; some who smoke and some who don't) with balanced values for related variables: age, sex, history of disease, etc.
2. A very strong correlation. For example, lung cancer is about 20 times more likely in smokers than non-smokers; the case for causation would be less convincing if it were only 2 times more likely.
3. Reproducability of results from many studies.

4. A dosage effect: more smoking correlates with more cancer.
5. Non-randomized Interventions: for example, show that people who quit smoking develop fewer cancers than those who continue.
6. Analogs: if we can't do randomized experiments on people, can we do them on animals? Or on cell tissue in a petri dish?
7. A theory: a known mechanism (such as the carcinogens in tobacco tar) for smoking causing cancer, and the lack of a theory for cancer causing smoking.

## Warning Sign I8: Believing Liars and Cheats


Cottingley Fairies

In the 1970s, Russell Targ and Harold E. Puthoff, two scientists at the Stanford Research Institute did experiments to evaluate the abilities of Uri Geller, and concluded that he had actual psychic powers. Later, magician James Randi showed how Geller performed his feats using standard conjuring tricks. Randi called Targ and Puthoff the "Laurel and Hardy" of psychic researchers. This may be harsh; their mistake seems to me not to be due to incompetence, but rather to trust: they couldn't believe that Geller would brazenly deceive people. Sixty years earlier a similar story played out when Arthur Conan Doyle, author of the Sherlock Holmes stories, became convinced of the legitimacy of several psychic mediums. It took another great magician of the day, Harry Houdini, to expose the tricks used by the mediums, but Houdini was unable to convince Doyle. Doyle also believed that the Cottingley Fairy photos (one shown at right) were legitimate evidence of fairies. Eventually, the girls who created the photos admitted that they were done with paper cutouts (Adobe Photoshop was not available in 1917). To a modern eye they look exactly like paper cutouts, but perhaps in 1917 people had less experience with photography, and with photographic fakes.

## Warning Sign I9: Being Too Clever

R.A. Fisher

Sir R. A. Fisher (1890-1962) was one of the greatest statisticians of all time, perhaps most noted for the idea of analysis of variance. But he sullied his reputation by arguing strongly that smoking does not cause cancer. He had some sensible arguments. First, he rightfully pointed out our Warning Sign I7, correlation is not causation. He was clever at coming up with alternative scenarios: perhaps lung cancer causes an irritation that the patient can feel long before it can be diagnosed, such that the irritation is alleviated by smoking. Or perhaps there is some unknown common cause that leads to both cancer and a tendency to smoke. Fisher was also correct in pointing out Warning Sign D1, lack of randomized trials: we can't randomly separate children at birth and force one group to smoke and the other not to. (Although we can do that with animal studies.) But he was wrong to be so dismissive of reproducible studies, in humans and animals, that showed a strong correlation, with clear medical theories explaining why smoking could cause cancer, and no good theories explaining the correlation any other way. He was wrong not to see that he may have been influenced by his own fondness for smoking a pipe, or by his libertarian objections to any interference with personal liberties, or by his employ as a consultant for the tobacco industry. Fisher died in 1962 of colon cancer (a disease that is 30% more prevalent in smokers than non-smokers). It is sad that the disease took Fisher's life, but it is a tragedy that Fisher's stuborness provided encouragement for the 5 million people a year who kill themselves through smoking. Note: since Fisher's death there have been some ingenious studies that largely get around the correlation problem, such as a study of 49 pairs of identical twins where one smokes and the other doesn't; the smokers were found to have more plaques on their carotid artery ($p$<0.1%). Also, the work of Judea Pearl, especially his book *Causality*, have advanced our understanding of the difference between correlation and causality.

## Conclusion

By now you should see that much can go wrong between the simple statement of "this result is significant at $p$=1%." and the conclusion about what that really means. As Darell Huff said, "it is easy to lie with statistics," but as Frederick Mosteller said, "it is easier to lie without them." By scrutinizing experiments against the checklist provided here, you have a better chance of separating truth from fiction.

## Bibliography


- The Canon, book by Natalie Angier, includes a chapter on statistics.
- Statistics for Experimenters, book by Box, Hunter and Hunter.
- Peter Donnelly: How juries are fooled by statistics, **video** of his TED talk.
- Electronic Statistics Textbook
- How to Lie with Statistics, classic 1954 book by Darrell Huff.
- Cartoon Guide to Statistics, book by Larry Gonick.
- Judgment under Uncertainty: Heuristics and Biases, book by Kahneman, Slovic, and Tversky.
- Innumeracy: Mathematical Illiteracy and Its Consequences, book by John Allen Paulos.
- Probability and Statistics EBook, by UCLA Statistics Depratment.
- What is a p-value anyway? 34 Stories to Help You Actually Understand Statistics by Andrew Vickers.
- Practical Statistics Simply Explained by Russell Langley, a cheap ($10) buut good Dover book.

- Common Statistical Mistakes, web presentation by Laura J. Simon.

## Acknowledgments

Thanks to Tim Josling, Thomas Lumley, Ravi Mohan, Seth Roberts, and Steve Simon for suggestions and corrections.

---

*Peter Norvig*