Evidence-Based Beliefs?

ADAM MORTON, UNIVERSITY OF BRITISH COLUMBIA

Progress

In ancient times, before some point in the second half of the nineteenth century, if you were uncertain how to investigate a topic, epistemologists—philosophers concerned with knowledge and rational belief—would be among the people you would first think of reading and consulting. They had played a large role in the early years of the scientific revolution, mediating the delicate tension between scientific discovery and traditional belief. The last such figure with this kind of influence was John Stuart Mill. But all that has changed. For at least the past hundred years, your first port of call would be a statistician.

There are several reasons for this. One is that the philosophers blew it. At first they were raising real issues about how to understand the physical world, and making helpful suggestions about how to achieve this. Some of these suggestions would seem bizarre now, but they were intelligently defended and usually fitted the science of the time. Then they got hung up on dramatic skeptical issues. How do we know the world is really there? How do we know that other people have minds? These are not stupid questions at all, and they link to contemporary issues in cognitive and social psychology. But they are not what your working physicist or medical researcher needs to ponder. There is also a flourishing discipline of philosophy of science, which both engages with the rest of philosophy and manages to say things of interest to working scientists. At its best it does, anyway. But, especially in recent decades, the philosophy of science has been concerned less and less with issues of method, and more with issues about the content of theories. Less how we should investigate and how much we should trust what we come up with, and more how we should understand what current views say.

At the same time, statistics has come into its own. There is a large, rich, and varied topic of statistical inference, concerned with drawing safe conclusions from varied or uncertain data. Very general and precise mathematical results can be applied to this. Moreover, some intellectual giants, such as the biologist/statistician Ronald Fisher and the brilliant original Abraham Wald, to name just two, showed us all that what might at first seem like a mere translation of methodological issues into mathematical terms is in fact really helpful for issues of what we should conclude, and how confident in our conclusions we should be. There is now no denying the relevance and irreplaceability of these ideas.

Personalities: Ronald Fisher introduced the idea of a randomized experiment, where a random selection of subjects is given the treatment in question, which is withheld from the rest. This goes a long way to neutralizing the effect of factors one had not anticipated. Ironically, he invented the theory that there might be a gene that predisposes people both to nicotine addiction and to lung cancer, so that quitting tobacco would make one less content without increasing one's life expectancy. (He smoked a pipe.) The irony is that randomized experiments are one of our better holds on telling correlation from causation. But in general separating these is very tricky and requires really high-powered statistics. His *Design of Experiments*, first written in the 1930s, has never been out of print.[1] He was also an influential biologist, one of the founders of population biology and one of the first to fuse Darwinian evolution with genetic theory. Abraham Wald invented sequential testing, which can drastically reduce the effort needed to reach a conclusion. Wald's brilliance is shown by advice he gave to the US Air Force in World War II. They had been reinforcing the areas of bombers that were most marked with bullet holes, but he said, "No: these are the planes that got back; you need to reinforce the unmarked areas, because that is where bullets struck the ones that didn't come back."

Problems

That's life: better things replace older things. Or so you might think, and in some ways you would certainly be right. But something is not right about the current situation. In this essay I describe some problems, make some suggestions about their causes, and tentatively explore a couple of remedies.

There is a paradox about the position of scientific evidence in contemporary life. On the one hand, the intrusion of sophisticated ideas into everyday concerns has never been greater. We routinely engage with cutting-edge research in ordinary decisions. We decide what cancer treatment to accept, depending on a medical report of the DNA typing of our tumors. We decide whether to spend a lot of money on a high-powered desktop computer, depending on the latest prognosis for Moore's law. We will soon have difficult public decisions about space exploration to make, depending on the latest thoughts about, for example, the technological problems of getting cargoes to other planets and the physiological problems of human spaceflight.

On the other hand, there is a widespread refusal among non-scientists to credit the force of scientific evidence. Large parts of the population, especially in the United States, simply ignore the evidence for global warming caused by human activity. Unsubstantiated rumors about the effects of vaccination have a hold on the public mind, in spite of the strong evidence against them. Public figures, dramatically in the United States but with many instances elsewhere, simply assert what they want and shrug off evidence to the contrary.

Low levels of public education, lower in some "advanced" countries than others, are part of the story. I believe that attitudes that make religious beliefs beyond reasonable criticism are also part of the story. But I shall discuss neither of these. Instead, I shall begin by discussing issues internal to science itself. Some of the things I say about underlying causes and possible remedies, in later sections, will also apply to the wider public.

Tainted Contexts

Drug companies needing government (e.g., FDA) approval of their products conduct large-scale trials to demonstrate their safety and efficacy. There are persistent stories of selectivity in the publication of these trials, so that inconvenient ones are ignored. This, outright fraud, is actually less worrying than some of the more subtle distortions associated with commercialized science. Peter Klamer, in his *Ordinarily Well: The Case for Antidepressants*, describes how antidepressants are given to large numbers of volunteers who are paid for their participation and classified, dosed, and monitored by reasonably trained personnel.[2] The volunteers are generally unemployed and living fairly marginal lives, and they have a motive to get themselves enrolled in the study. The people administering the trial need to evaluate a certain proportion as suitably depressed, though the evaluation of their reactions to the drugs or placebos is probably honest.

The most worrying result of this situation is that the pool of subjects is most likely significantly different from that of the patients who would receive the drug. They will contain a greater proportion of people who are depressed because their lives are miserable rather than because of some malfunction in their thinking or brain chemistry. And they have different motives for responding to questions in a way that will lead to a diagnosis. So the transfer of the results of the study to the larger class of depressed patients is pretty problematic. The formal conduct of the experiment is correct, good models are superficially emulated, but it is undermined by the motivated way in which the details are chosen.

There is a connection here with a theme in the writings of the philosopher of science Nancy Cartwright. Cartwright argues that properly conducted experiments are a fine way of establishing causal connections in the experimental context, but that there is a very serious question when we can extend these connections to nonexperimental situations. This is a particularly important issue with medical results.

Another personality: Cartwright is one of the most eminent contemporary philosophers of science. She is best known for defending, in her *How the Laws of Physics Lie* and in later books, the view that what we call laws of nature are patchy and exception-ridden generalizations smoothing out the real, more complex ways in which one event causes another.[3] In more recent work she has argued that we have a variety of concepts of causation, which need a variety of methods. (So there is a connection with Fisher on smoking.) Some of her conclusions here might be summed up as: it is fine to talk about evidence-based medicine, but wouldn't it be a good idea first to understand what evidence is?

Nonreplication

It should be no surprise that many experiments give different results when later repeated in different laboratories. There is always noise in the data, and there are always variations between different samples even when they are drawn at random from a homogeneous population. For the same reasons, we should expect that sometimes the effect is real and is detected in an initial experiment but missed in an attempted replication. However, in recent years in several experimental sciences, a disturbing proportion of experiments have proved not to replicate. One analysis estimates that more than half of the reported results in psychology will not replicate because the effect they describe does not exist.[4]

Publication Bias

It is easier to get something published, and thus advance one's career, if it presents a new result rather than confirming a previously announced one. Journal editors are more likely to give it space, and one's colleagues are more likely to be impressed. But this tends against replication and the dissemination of confirmatory studies. The same motives can be found in the "desk drawer problem": researchers may not even try to publish results that may be met with a yawn. More subtly, a trial or preliminary study that suggests that a fuller or more careful experiment would not break new ground may often not be followed by such a fuller or more careful version.

Journal editors are more favorably disposed to reports of experiments with larger sample sizes, more safeguards, and better design than those that they replicate. These can seem more significant. If the original experiment was clearly inadequate, then the case for an improved replication is not difficult to make. But when the original was adequate, the case for a better replication is harder to pitch. The situation is complicated by the use of conventional significance levels: if the original would be standardly taken as making a case for its conclusion, then it may be hard to see that a replication making a better case is adding to what we already know. Moreover, the effort and expense of a larger or more elaborate experiment may not be easy to justify given that the result may not turn out to be more significant or give more definite conclusions. It is an aversion to the unglamorous.

Confirmation Conflation

Some statistical tests, notably Fisherian significance tests, which estimate how likely it is that the observed result happened by chance, examine the existence of a phenomenon. Others, especially some likelihood-based tests, which compare estimates of the probability of given data conditional on each of two possible explanations, examine the comparative support for two hypotheses. The question for the first is "is anything going on here?" and the question for the second is "which of these best accounts for the evidence?" These can be confused. In the most grotesque confusion we have a null hypothesis that nothing unusual has happened and an explanatory hypothesis that they are not random but the consequences of a specified mechanism. We run the experiment, compare the data to the null hypothesis, and conclude that (there is a good chance that) they are not purely random. So far so good, but then we infer that the explanatory hypothesis is true! The craziness of this emerges when we consider that the explanatory hypothesis played no role in the test. Any other hypothesis could have been substituted, and its truth could equally well have been established. It is as if we consider the null hypothesis that the coin is fair against an alternative that it is being biased by telepathic influences from alpha centauri, toss it getting fourteen heads and six tails, and then since that is significantly different from the behavior of a fair coin conclude that we have confirmed extraterrestrial telepathy. But procedures like this are routine in some scientific disciplines.[5]

There is an interesting and important sociological dimension to this. There have been many reports in the press of the troubles of experimental science, and some reassurances that things are not as bad as they might seem.[6] One factor that is beginning to be studied is whether press and other reports of inevitable scientific disagreement tend to adopt an even-handedness that leads readers to think that the balance of evidence is less definitive than it is.[7] A recent study describing weakness in the evidence that flossing reduces tooth decay has received much gleeful publicity in the media, often with the theme "so perhaps flossing does you no good after all." But what the study was pointing out was that while we have evidence in favor of flossing, it does not consist in controlled randomized experiments. Again, we see Fisherian issues at play.

Diagnosis

Our official practices are better than they ever have been, and our knowledge of how these practices should work is better than it ever has been, but things often go wrong. I think there are two central reasons. They both come down, one way or another, to the fact that statistical inference is a mathematically very demanding topic.

Cookbooks

Most working scientists, even in disciplines that require a mathematical background, are not familiar with the details of statistical inference. And if they try getting up to date on this, they soon find that it is hard, confusing, and in some ways unlike other parts of mathematics. (The standard notation, for one thing, has peculiar quirks.) And to their surprise they find that although there are agreements on many central issues, the topic is also full of controversies between rival schools where extremely subtle results are tossed around. Many of these results are graspable with enough patience and mathematical background, but their relevance to the issues under dispute is clear only to those who have spent years fighting about them. To make things worse, statisticians tend to agree that many issues about the design of experiments will be clear to anyone indoctrinated into their profession but hard to explain to any outsider.

Things get even harder if we consider Cartwright's problem. For deciding whether an experimental result holds in a nonexperimental context is a very tricky business. So what is the poor researcher to do? She doesn't want to tackle the statistics from first principles. She finds statisticians hard to talk to. And the controversies between statisticians leave her bewildered. The usual solution is that she applies some rule of thumb or standard computer program that is current in her discipline. Often these are inappropriate for the particular case at hand, or applying them in a careful and fitting way would require just the very sophistication that is lacking. In many academic departments there is one colleague who is taken by others as an authority on the things, though they rarely operate at a really professional level. So if others are following some routine and getting away with it, you will too. You will get papers published and you will get your promotions.

The results will often be flawed, increasing the chance that they will not replicate. And sometimes seen from a sophisticated point of view, or just the point of view of a different discipline where practices are different (so they go for different oversimplifications), they will emerge as simply grotesque—thus the mechanical misuse of significance levels in psychology and other disciplines.

Authority

If you don't have the specialized mathematical training and the years of experience that a professional statistician has, you have to rely on a computer program, authorities in your field, what journal editors are happy with, or what a tame statistician tells you. (These can conflict; the statistics department at your university may be at odds with the schools of thought with which your discipline's top journal is cozy.) These are sources of authority, and authority plays a big role in scientific practice, often but not always for the good. So just as in your graduate school education the route to discovering truth and the route to professional success were closely associated, in your practice as a member of a discipline what you do is equal parts pleasing the influential figures and asserting what seems to you to be made probable by the evidence. It would require almost superhuman self-knowledge and reflection to separate these two in your own case.

The result is conformity and deference. Not at all necessarily a bad thing, when those in authority really do know best. But it is becoming pretty clear that they often do not. One sign that something is amiss is the hodgepodge of views about causal inference, the lore of telling correlation from causation. Everyone who has taken an introductory statistics course has learned the mantra "correlation does not show causation." Additional information and additional technique are needed. But which additional information depends on which additional technique, and schools of thought about this are as diverse as small Protestant sects. Unless you are going to be a deep and original methodologist, which is best if you are brilliant and established, the route to survival is to knuckle down and do what the powers in your discipline do.

Remedies

My theme is that methods are good—we have never before had such powerful tools at our disposal—yet the results are often bad—irresponsible studies are routinely published, experiments don't replicate, and whole areas of research have an air of disrepute. What is to be done?

Real Epistemology

Nothing is going to displace statistical inference as the ultimate arbiter of the force of evidence. And the disputes between schools of statisticians are healthy science, necessary for us to make progress on these fundamental—and philosophical—issues. The aim has to be to take away its remote magisterial air. One way of achieving this would be for courses in statistics and experimental method designed for graduate students in particular disciplines to focus less on what is standard procedure and more on the problems that standard procedure addresses.

Courses along these lines would be more confusing, in the typical manner of philosophy as opposed to science. Confusion might be a good thing.

A related measure would be for the philosophers to get their act together. Their activity could be presented as wrestling with problems of real-world belief formation rather than as elegant approaches to skeptical conundra. (I write "presented as" because in fact the rarefied parts of the subject do engage with real belief formation. But this is not evident to anyone much behind the cutting edge.) The issues that would emerge would be more closely related to the issues of appropriate statistical procedure. The philosophers ought to learn some statistics and the statisticians ought to learn to take epistemology seriously.

Being Explicit about Authority

Many things are less powerful if they are said out loud. Scientific authority is generally a necessary and beneficent thing. Unstated implicit scientific authority is what can do harm, and encourage groveling and imitation. So the social structure of each discipline should be transparent to everyone in it and stated explicitly. The authority of statisticians, also. There is no substitute for relying on the knowledge and procedures of those who are rightly known as the leaders of their field. But questions should always be in the air about which these are and how they have earned it.

Participation

You don't understand the authority structures of institutions you are not part of. You don't understand other people's misconceptions unless you work with them and talk to them. You have much less influence on what people do if you do not share projects with them. There there are three tribes this affair: the scientists, the statisticians, and the general public. (Scientists in one discipline have a half-affinity to scientists in a very different discipline: they sort of know what the others are dealing with, but many important details are not available to them.) Ideally, every citizen would have a scientific project that she keeps up with, knows the research on, and where she can tell reputable from sloppy work. Ideally, every scientist would have some corner of statistics that she keeps up with, and some topic of public concern that she relates to her work. Ideally, every statistician would think about the use that practicing scientists make of her work. Ideally: but it is not too much to ask that we develop institutions that encourage these practices.

There is a common theme to these suggestions: stating what the rules and practices are, including the rules and practices of those you connect with only indirectly, and making a project out of improving them. In science as in any other society, that is what the balance between accomplishing individual aims and achieving shared goals requires.

Notes

1. Ronald Fisher, *Design of Experiments* (Edinburgh: Oliver & Boyd, 1935).

2. Peter Klamer, *Ordinarily Well: The Case for Antidepressants* (New York: Farrar, Straus & Giroux, 2016).

3. Nancy Cartwright, *How the Laws of Physics Lie* (Oxford: Clarendon, 1983).

4. See Open Science Collaboration, "Estimating the Reproducibility of Psychological Science," *Science* 349, no. 6251 (2015): http://science.sciencemag.org/content/349/6251/aac4716. This is an instance of metastatistics, where statistical techniques are used to assess the reliability of statistical practice.

5. The most telling accusations here come from the writings of eminent psychologists attacking the corner-cutting of their colleagues. See Gerd Gigerenzer, "Mindless Statistics," *Journal of Socio-Economics* 33 (2004): 587–606; and Richard Nisbett, "The Crusade against Multiple Regression Analysis, *Edge*, January 21, 2016, http://edge.org/conversation/richard_nisbett-the-crusade-against-multiple-regression-analysis. Two very major figures horrified at their colleagues' practices.

6. See, e.g., George Johnson, "New Truths That Only One Can See," *New York Times*, January 20, 2014; Benedict Carey, "New Critique Sees Flaws in Landmark Analysis of Psychology Studies" *New York Times*, March 3, 2016; Christie Aschwanden, "Science Isn't Broken," *FiveThirtyEight* (blog), August 19, 2015, https://fivethirtyeight.com/features/science-isnt-broken/.

7. See Derek J. Koehler, "Why People Are Confused about What Experts Really Think," *New York Times*, February 12, 2016.