Error and the Growth of Experimental Knowledge

by Deborah G. Mayo

Science and Its Conceptual Foundations series
University of Chicago Press, 1996

We Have Ways of Making You Talk, or, Long Live Peircism-Popperism-Neyman-Pearson Thought!

After I'd bungled teaching it enough times to have an idea of what I was doing, one of the first things students in my introductory physics classes learned (or anyway were taught), and which I kept hammering at all semester, was error analysis: estimating the uncertainty in measurements, propagating errors from measured quantities into calculated ones, and some very quick and dirty significance tests, tests for whether or not two numbers agree, within their associated margins of error. I did this for purely pragmatic reasons: it seemed like one of the most useful things we were supposed to teach, and also one of the few areas where what I did had any discernible effect on what they learnt. Now that I've read Mayo's book, I'll be able to offer another excuse to my students the next time I teach error analysis, namely, that it's how science really works.

I exaggerate her conclusion slightly, but only slightly. Mayo is a dues-paying philosopher of science (literally, it seems), and like most of the breed these days is largely concerned with questions of method and justification, of "ampliative inference" (C. S. Peirce) or "non-demonstrative inference" (Bertrand Russell). Put bluntly and concretely: why, since neither can be deduced rigorously from unquestionable premises, should we put more trust in David Grinspoon's ideas about Venus than in those of Immanuel Velikovsky? A nice answer would be something like, "because good scientific theories are arrived at by employing thus-and-such a method, which infallibly leads to the truth, for the following self-evident reasons." A nice answer, but not one which is seriously entertained by anyone these days, apart from some professors of sociology and literature moonlighting in the construction of straw men. In the real world, science is alas fallible, subject to constant correction, and very messy. Still, mess and all, we somehow or other come up with reliable, codified knowledge about the world, and it would be nice to know how the trick is turned: not only would it satisfy curiosity ("the most agreeable of all vices" --- Nietzsche), and help silence such people as do, in fact, prefer Velikovsky to Grinspoon, but it might lead us to better ways of turning the trick. Asking scientists themselves is nearly useless: you'll almost certainly just get a recital of whichever school of methodology we happened to blunder into in college, or impatience at asking silly questions and keeping us from the lab. If this vice is to be indulged in, someone other than scientists will have to do it: namely, the methodologists.

That they have been less than outstandingly successful is not exactly secret. Thus the biologist Peter Medawar, writing on Induction and Intuition in Scientific Thought: "Most scientists receive no tuition in scientific method, but those who have been instructed perform no better as scientists than those who have not. Of what other branch of learning can it be said that it gives its proficients no advantage; that it need not be taught or, if taught, need not be learned?" Still, they have made some progress: at least since William Whewell's 1840 Philosophy of the Inductive Sciences, those of them who are (as the saying goes) sharper than a sack of wet mice have realized that it's much easier to get rid of wrong notions than to find correct ones, if the latter is possible at all. In our own time, Medawar's friend Karl Popper achieved (fully deserved) eminence by tenacious insistence on the importance of this point, becoming a sort of Lenin of the philosophy of science. Instead of conferring patents of epistemic nobility, lawdoms and theoryhoods, on certain hypotheses, Popper hauled them all before an Anglo-Austrian Tribunal of Revolutionary Empirical Justice. The procedure of the court was as follows: the accused was blindfolded, and the magistrates then formed a firing squad, shooting at it with every piece of possibly-refuting observational evidence they could find. Conjectures who refused to present themselves might lead harmless lives as metaphysics without scientific aspirations; conjectures detected peaking out from under the blindfold, so as to dodge the Tribunal's attempts at refutation, were declared pseudo-scientific and exiled from the Open Society of
Science. Our best scientific theories, those Stakhanovites of knowledge, consisted of those conjectures which had survived harsh and repeated sessions before the Tribunal, demonstrated their loyalty to the Open Society by appearing before it again and again and offering the largest target to refutation that they could, and so retained their place in the revolutionary vanguard until they succumbed, or were displaced by another conjecture with even greater zeal for the Great Purge. (The whole affair was very reminiscent of _The Golden Bough_, though I don't know if Popper ever read it; also of Nietzsche's quip that "it is not the least charm of a hypothesis that it is refutable.") As Popper famously said, better our hypotheses die for our errors than ourselves... It's an answer with nice, clean lines, and makes lots of sense to the scientist-at-the-bench, like Medawar. Alas, the Revolution runs into trouble on several fronts, for instance statistics. Suppose I tell you that a certain slot machine will pay out money 99% of the time. Being credulous, unnaturally patient, and abundantly supplied with coins, you play it 10,000 times and find that it pays out only twice. This is sufficient for you to tell me to get stuffed, if not to sue, and one would think that it would be enough for the Tribunal to shoot my poor conjecture dead, but actually it escapes unharmed. The problem for Uncle Karl is that getting two successes in ten thousand trials is possible given my assertion, and the Tribunal is only authorized to eliminate conjectures in actual contradiction to the facts, as "no mammals lay eggs" is contradicted by the platypus. Popper realized this, and worried about it, eventually saying that we just have to make "risky decisions" about when to reject statistical hypotheses. But the challenges facing the Tribunal in the execution of its duty mount: another "risky decision" is required, about what ammunition the firing squad can legitimately use, i.e., about what evidence will be accepted when we see whether or not a hypothesis stands up. (The number of times my students have apparently refuted physical laws gives me great sympathy for the European naturalists who refused to accept reports of the platypus's peculiarities for decades.) Then there is the problem of conjectural conspiracy: an isolated hypothesis almost never leads to anything we can test observationally; it is only in combination with "auxiliary" hypotheses, sometimes very many of them indeed, that is gives us actionable predictions. But then if a prediction proves false, all we learn is that at least one of our hypotheses is wrong, not which ones are the saboteurs. So far as deductive rectitude is concerned, we are free to frame whichever auxiliaries we like least, and save our favorite hypothesis from execution at the hands of the Tribunal. The Tribunal even, for all its appearance of salutary rigor, lets far too many suspects go: _every_ conjecture which is compatible with the evidence. These last two problems, respectively those of Quine-Duhem and of methodological underdetermination, are so severe that they form the core of the (intellectually respectable) argument for the counter-revolutionary deviation of scientific relativism. (The argument throttles itself neatly, but that's a subject for another essay.) Yet in ordinary life, never mind science, we evade these problems --- those of testing statistical hypotheses, of selecting evidence, of Quine-Duhem, of methodological underdetermination --- every time we change a light-bulb, so something has clearly gone very wrong here (as, in revolutions, things are wont to do).

Mayo, playing the Jacobin or Bolshevikk to Popper's Girondin or Cadet, thinks she knows what the problem is: for all his can't-make-an-omelette-without-breaking-eggs rhetoric, Popper is entirely too soft on conjectures.

Although Popper's work is full of exhortations to put hypotheses through the wringer, to make them "suffer in our stead in the struggle for the survival of the fittest," the tests Popper sets out are white-glove affairs of logical analysis. If anomalies are approached with white gloves, it is little wonder that they seem to tell us only that there is an error somewhere and that they are silent about its source. We have to become shrewd inquisitors of errors, interact with them, simulate them (with models and computers), amplify them: we have to learn to make them talk. [p. 4, reference omitted]

Fortunately, scientists have not only devoted much effort to making errors talk, they have even developed a theory of inquisition, in the form of mathematical statistics, especially the theory of statistical inference worked out by Jerzy Neyman and Egon Pearson in the 1930s. Mayo's mission is largely to show how this very standard mathematical statistics justifies a very large class of scientific inferences, those concerned with "experimental knowledge," and to suggest that the rest of our business can be justified on similar grounds. Statistics becomes a kind of applied methodology, as well as the "continuation of experiment by other means."
Mayo's key notion is that of a **severe test** of a hypothesis, one with "an overwhelmingly good chance of revealing the presence of a specific error, if it exists --- but not otherwise" (p. 7). More formally (when we can be this formal), the severity of a passing result is the probability that, if the hypothesis is false, our test would have given results which match the hypothesis less well than the ones we actually got do, taking the hypothesis, the evidence used in the test, and the way of calculating fit between hypothesis and evidence to be fixed. [Semi-technical note containing an embarrassing confession.] If a severe test does not turn up the error it looks for, it's good grounds for thinking that the error is absent. By putting our hypotheses through a battery of severe tests, screening them for the members of our "error repertoire," our "canonial models of error," we can come to have considerable confidence that they are *not* mistaken in those respects. Instead of a method for infallibly or even reliably finding trutths, we have a host of methods for reliably finding errors: which turns out to be good enough.

Experimental inquiry, for Mayo, consist of breaking down the question at hand into a series of small bits, each of which is relatively easily subjected to severe tests for error, or (depending on how you look at it) is itself a severe probe for a certain error. In doing this we construct a "hierarchy of models" (an idea of Patrick Suppes's, here greatly elaborated). In particular, we need *data models*, models of how the data are collected and massaged. "Error" here, as throughout Mayo's work, must be understood in a rather catholic sense: any deviation from the conditions we assumed in our reasoning about what the experimental outcomes should be. If we guess that a certain effect (the bending of spoons, let us say) is due to a certain cause (e.g., the psychic powers of Mr. Uri Geller), it is not enough that spoons bend reliably in his presence: we must also rule out other mechanisms which would produce the same effect (Mr. Geller's bending the spoons with his hands while we're not looking, his substituting pre-bent spoons for unbent ones ditto, etc., through material for several lawsuits for libel). But this solves the Quine-Duhem problem.

In fact, it gets better. Recall that methodological underdetermination (which goes by the apt name of MUD in *Error*) is the worry that no amount or quality of evidence will suffice to pick out one theory as the best, because there are always indefinitely many others which are in equal accord with that evidence, or, to use older language, equally well save the phenomena. But saving the phenomena is *not* the same as being subjected to a severe test: and, says Mayo, the point is severe testing. While I'm mostly persuaded by this argument, I'm less sanguine than Mayo is about our ability to *always* find experimental tests which will let us discriminate between two hypotheses. I'm fully persuaded that this kind of testing really does underwrite our knowledge of *phenomena*, of (in Nancy Cartwright's phrase) "nature's capacities and their measurement," and Mayo herself insists on the importance of experimental knowledge in just this sense (e.g., the remarks on "asking the wrong question," pp. 188--9). I'm less persuaded that we can usually or even often make justified inferences from this "formal" sort of experimental knowledge, knowledge of the distribution of experimental outcomes, to "substantive" statements about objects, processes and the like (e.g., from the experimental success of quantum mechanics to wave-functions). As an unreconstructed (undeconstructed?!) scientific realist, I make such inferences, and would like them to be justified, but find myself left hanging. (Mayo is currently working on the connection between experimental knowledge, fairly low in the hierarchy of models, and the higher-level theories philosophers of science have more traditionally fretted over, i.e., points more or less like this one.)

Distributions of experimental outcomes, then, are the key objects for Mayo's tests, especially the standard Neyman-Pearson statistical tests. The kind of probabilities Mayo, and Neyman and Pearson, use are probabilities of various things happening: meaning that the probability of a certain result, p(A), is the proportion of times A occurs in many repetitions of the experiment, its frequency. This is a very familiar sense of probability; it's the one we invoke when we say that a fair coin has a 50% probability of coming up heads, that the chance of getting three sixes with fair (six-sided!) dice is 1 in 216, that a certain laboratory procedure will make an indicator chemical change from red to blue 95% of the time when a toxin is present. Or, more to the present point: "the hypothesis is significant at the five percent level" means "the hypothesis passed the test, and the probability of its doing so, if it were false, is no more than five percent," which means "if the hypothesis is false, and we repeated this experiment many times, we would expect to get results inside our passing range no more than five percent of the time."

This interpretation of probability, the "frequentist" interpretation, is not the only one however. Ever since its origins in the seventeenth century, if we are to believe its historians, mathematical probability has
oscillated, not to say equivocated, between two interpretations, between saying how often a given kind of event happens, and saying how much credence we should give a given assertion. Now, this is the sort of philosophical question --- viz., what the hell is a probability anyway? --- which scientists are normally none the worse for ignoring, and normally blithely ignore. But maybe once every hundred years these questions actually affect the course of research, philosophy really does make a difference: the existence of atoms was such a question at the beginning of the century, and the nature of probability is one today. To see why, and why Mayo spends much of her book chastising the opponents of the frequentist interpretation, requires a little explanation.

Modern believers in subjective probability are called Bayesians, after the Rev. Mr. Thomas Bayes, who in 1763 posthumously published a theorem about the calculation of conditional probabilities, which runs as follows. Suppose we have two classes of events, A and B, and we know the following probabilities: p(A), the probability of A, all else being equal; p(B), the probability of B, likewise; and p(B|A), the probability of B given A. Then we can calculate p(A|B), the probability of A given B: it's p(B|A)p(A)/p(B). The theorem itself is beyond dispute, being an easy consequence of the definition of a conditional probability, with many useful applications, the classical one being diagnostic testing. The uses to which it has been put are, however, as peculiar as those of any mathematical theorem, even Gödel's.

In particular, if you think of probabilities as degrees-of-belief, it is tempting, maybe even necessary, to regard Bayes's theorem as a rule for assessing the evidential support of beliefs. For instance, let A be "Mr. Geller is psychic" and B be "this spoon will bend without the application of physical force." Once we've assigned p(A), p(B), and p(B|A), we can calculate just how much more we ought to believe in Geller's psychic powers after seeing him bend a spoon without visibly doing so. p(A) and p(B) and sometimes even p(B|A) are, in this view, all reflections of our subjective beliefs, before we examine the evidence. They are called the "prior probabilities," or even just the "priors." The prize, p(A|B), is the "posterior," and regarded as the weight we should give to a hypothesis (A) on the strength of a given piece of evidence (B). As I said, it's hard to avoid this interpretation if you think of probabilities as degrees-of-belief, and there is a large, outspoken and able school of methodologists and statisticians who insist that this is the way of thinking about probability, scientific inference, and indeed rationality in general: the Bayesian Way.

Looked at from a vantage-point along that Way, Neyman-Pearson hypothesis testing is arrant nonsense, involving all manner of irrelevant considerations, when all you need is the posterior. For those of us taking the frequentist (or, as Mayo prefers, error-statistical) perspective, Bayesians want to quantify the unquantifiable and proscribe inferential tools that scientific practice shows are most useful, and are forced to give precise values to perfectly ridiculous quantities, like the probability of a getting a certain experimental result if all the hypotheses we can dream up are wrong. For us, to assign a probability to a hypothesis might make sense (in Peirce's words) "if universes were as plenty as blackberries, if we could put a quantity of them in a bag, shake them well up, draw out a sample and examine them" (Collected Works 2.684, quoted p. 78); as it is, hypotheses are either true or false, a condition quite lacking in gradations. Bayesians not only assign such probabilities, they do so a priori, condensing their prejudices into real numbers between 0 and 1 inclusive; two Bayesians cannot meet without smiling at each other's priors. True, they can show that, in the limit of presenting an infinite amount of (consistent) evidence, the priors "wash out" (provided they're "non-extreme," not 0 or 1 to start with); but it has also been shown that, "for any body of evidence there are prior probabilities in a hypothesis H that, while nonextreme, will result in the two scientists having posterior probabilities in H that differ by as much as one wants" (p. 84n, Mayo's emphasis). This is discouraging, to say the least, and accords very poorly with the way that scientists actually do come to agree, very quickly, on the value and implications of pieces of evidence. Bayesian reconstructions of episodes in the history of science, Mayo says, are on a level with claiming that Leonardo da Vinci painted by numbers since, after all, there's some paint-by-numbers kit which will match any painting you please.

Mayo will have nothing to do with painting by numbers, and wants to trash all the kits she runs across. These do not just litter the Bayesian Way; the whole attempt to find "evidential relation" measures, which will supposedly quantify how much support a given body of evidence provides for a given hypothesis, fall into the dumpster as well. The idea behind them, that the relation between evidence and hypothesis is some kind of a fraction of a deductive implication, can now I think be safely set aside as a nice idea which just
doesn't work. (This is a pity; it is easy to program.) It should be said, as Mayo does, that the severity of a test is not an evidential relation measure, rather is a property of the test, telling us how reliably it picks out a kind of mistake --- that it misses it once every hundred tries, or once every other try, or never. (If a hypothesis passes a test on a certain body of evidence with severity 1, it does not mean that the evidence implies the hypothesis, for instance.) Also on the list of science-by-numbers kits to be thrown out are some abuses of Neyman-Pearson tests, the kind of unthinking applications of them that led a physicist of my acquaintance to speak sarcastically of "statistical hypothesis testing, that substitute for thought." Some of these Mayo lays (perhaps unjustly) at Neyman's feet, exonerating Pearson; she shows that none of them are necessitated by a proper understanding of the theory of testing.

In the next to last chapter Mayo tries her hand at one of American philosophy's perennial amusements, the game of Peirce Knew It All Along. (If, as Whitehead said, European thought is a series of footnotes to Plato, American thought is a series of footnotes to Peirce --- and Jonathan Edwards, worse luck.) Usually this is a mere demonstration of cleverness, like coining words from the names of opponents, or improving on the proof that if $1+1=3$, then Bertrand Russell was the Pope. But in this case it seems that Mayo is really on to something. It is sometimes forgotten that Peirce was by training an experimental scientist, was employed as an experimental physicist for years, and as such lived and breathed error analysis. His opposition to subjective probabilities and paint-by-numbers inductivism is plain. For him "induction" meant the experimental testing of hypotheses; the probabilities employed in induction are the probabilities of inductive procedures leading to correct answers:

The theory here proposed does not assign any probability to the inductive or hypothetic conclusion, in the sense of undertaking to say how frequently that conclusion would be found true. It does not propose to look through all the possible universes, and say in what proportion of them a certain uniformity occurs; such a proceeding, were it possible, would be quite idle. The theory here presented only says how frequently, in this universe, the special form of induction or hypothesis would lead us right. The probability given by this theory is in every way different --- in meaning, numerical value, and form --- from that of those who would apply to ampliative inference the doctrine of inverse chances [i.e., Bayes's theorem]. [2.748, quoted p. 414]

But severity, and the related error probabilities, say, exactly, how often a particular procedure of "ampliative inference" will "lead us right." Most of the rest of Mayo's approach is hinted at in volumes 2 and 7 of the Collected Works as well --- the hierarchy of models of experiment (well, a hint of a hint anyway: "The secret of the business lies in the caution which breaks a hypothesis up into its smallest logical components, and only risks one of them at a time," 7.220, quoted p. 434, Mayo's emphasis), the need for canonical models of error and an error repertoire, for modeling the generation of data, the self-correcting nature of induction --- not so much that truth will prevail, as that errors will amplify and come out. In the immortal words of Piet Hein:

*The road to wisdom? --- Well, it's plain and simple to express:*

Err
and err
and err again
but less
and less
and less.

Then, too, there is the interesting, and I think absolutely correct, view of the purpose and utility of a theory of experiment: "It changes fortuitous events, which may take weeks or may take many decennia, into an operation governed by intelligence, which will be finished within a month" (7.78, quoted p. 434). This is of a piece with the general function of intellectual traditions. Genius can, perhaps, get by on its wits, make things up from scratch, etc. Intellect serves the rest of us, by codifying, by setting up standards and procedures which can be followed with only (as a friend once happily put it) "a mediocum of intelligence," so that what might have taken genius can be (at least partially) achieved through the application of rules. Among those rules, "normal tests" or "standard tests" --- tests which have proved to be reliable detectors of specific errors --- take a special place. Traditions of inquiry which incorporate and use a family of normal
tests may fail to produce reliable knowledge, but those which don't can hardly hope even to produce interesting mistakes.

There have been earlier attempts to ground the philosophy of science on statistical theory, even the Neyman-Pearson theory, most notably Braithwaite's *Scientific Explanation*. Mayo's book is superior to them: at least as brilliant, and for once doing the jobs which need doing. By argument and by example (e.g., the two very detailed case studies of Perrin's experiments on Brownian motion, and the observations of the solar eclipse of 1919, both testing and --- as it happens --- confirming theories of Einstein's) she really does show how important methodological problems are solved in scientific practice. Her writing is less than stellar (the passage I quoted about making errors talk is the stylistic high point of the book), but entirely adequate to the task, which is much more than can be said for most philosophical books, much less those on the philosophy of statistics. There is mathematics, but it's fairly simple and self-contained; one needn't worry about being suddenly confronted with a proof of the Neyman-Pearson Lemma, or even of the Law of Large Numbers. Mayo succeeds in everything important she sets out to do; she may even have succeeded, in her long discussions of Kuhn (in chs. 2 and 4) in defanging him, but I frankly couldn't work up enough interest in her interpretation of Kuhn's interpretation of Popper (sometimes, her interpretation of other people's interpretations of Kuhn's interpretation of Popper) to see if she really succeeds in turning Kuhn's sociological descriptions into methodological prescriptions. (There is very little about the social aspects of science in this book; oddly, it does not feel like a flaw.)

Aside from my usual querulousness about style (and it's not fair to hold not writing as well as Russell or Dennett or Quine against a philosopher who actually does write decently), I have only two substantial problems with Mayo's ideas; or perhaps I just wish she'd pushed them further here than she did. First, they do not seem to distinguish scientific knowledge --- at least not experimental knowledge --- from technological knowledge, or even from artisanal know-how. Second, they leave me puzzled about how science got on before statistics.

Experimental knowledge (taking first things first) is, for Mayo, pretty much knowing what happens in certain circumstances --- knowing how to reliably produce certain effects. But this doesn't serve to distinguish between, say, a condensed matter physicist and a metallurgical engineer, or even between them and a medieval blacksmith from Damascus, who may all be concerned with the same process, and all know that if you take iron strips and hammer them together between repeated forgings you get a stronger metal than by just casting the same amount of the same iron in the same final shape. It is far from clear to me that her demarcation criterion --- "What makes an empirical inquiry scientific is that it can and does allow learning from normal tests, that it accomplishes one or more tasks of normal testing reliably" (p. 36), --- does the job; certainly not as between science and engineering. Indeed, Mayo makes a point of noting that "arguing from error" is part of everyday life. I'm quite sympathetic to the idea that the distinction between what we call "science" and other sorts of reliable knowledge (or, if you like, other reliable practices of inquiry) does not reflect any deep methodological divide, but, say, is one of subject-matter, or even of the adventitious history of English usage; but then that same usage makes it misleading to call the things on one side of the methodological divide "scientific" and the others "unscientific."

Which leads to the other worry: there was lots of good science long before there were statistical tests; Galileo had reliable experimental knowledge if anyone did, but error analysis really began two centuries after his time. (If we allow engineers and artisans to have experimental knowledge within the meaning of the act, we can push this back essentially as far as we please.) If experimental knowledge is reached through severe tests, and the experimenters knew not statistical inference, then the apparatus of that theory isn't necessary to formulating severe tests. But how then do we know that they're really severe? Presumably in the same way in which we mundanely argue from error, more or less intuitively. If this intuition led us in our wanderings from the Goshen of superstition to the Canaan of statistical inference, it would be nice to understand it, and why we are blessed with it when (say) rats are not (are they?), and why it is not or was not applied to some subjects. (It would be fascinating to re-examine intellectual and technological history as the evolution of error-probes; probably also pretty depressing, at least on the intellectual side.) Let us put such quibbles aside. Anyone with a serious interest in how science works ought to read this. It will even be useful to scientists: for a work on the philosophy of science, this places it above rubies.
Disclaimer: Prof. Mayo was kind enough to look over this review, and save me from at least one really gross mistake; but I have no stake in the success of *Error and the Growth of Experimental Knowledge*, and she shouldn't be held responsible for any goofs in which I have persisted through mule-headedness.

xvi+493pp., frontispiece pencil sketch by the author of Egon Pearson, black and white graphs, digressive footnotes, bibliography, analytical index

*Philosophy of Science / Probability and Statistics*


With thanks to Rob Haslinger for turns of phrase; Tony Lin and Erik van Niemwegen for arguments about statistics; and my students in intro physics.

11--14 September 1998

Typo fix 31 July 2006, thanks to Dave Kane

Link fix 22 October 2007, thanks to Ed Johnston