Record: 1

INSTRUMENTAL PROBABILITY. By: Glymour, Clark. Monist, Apr2001, Vol. 84 Issue 2, p284, 17p; (*AN 4562569*)

Database: Academic Search Premier

INSTRUMENTAL PROBABILITY

1. The Science of Nothing

The claims of science and the claims of probability combine in two ways. In one, probability is part of the content of science, as in statistical mechanics and quantum theory and an enormous range of "models" developed in applied statistics. In the other, probability is the tool used to explain and to justify methods of inference from records of observations, as in every science from psychiatry to physics. These intimacies between science and probability are logical sports, for while we think science aims to say what happens, what has happened, what will happen, what would happen if other things were to happen, what could and could not happen, what will nearly happen, or what will approximate what will happen, probability claims say none of these things, or at least none of them about the phenomena with which science is concerned. On that, at least, almost all philosophical interpreters of probability since DeMoivre agree, with whatever reluctance. Consider some examples.

According to Hume and to many moderns, probability is a measure of opinion, of credence or degree of belief. But measures of opinion are not statements about what happens in nature or society, what has happened, what will happen, what would happen if other things were to happen, etc. They are, at most, statements about the mental state of the opinion-holder, or about what the opinion-holder will do if offered wagers. To press the obvious, from "Hume's degree of belief that the sun will rise tomorrow is .9999" it does not follow that the sun will rise tomorrow, or that it will not. Thomas Bayes, along with a smaller group of our contemporaries, defined probability as a norm of belief, a measure of what one ought to believe. Norms of belief are as remote from empirical claims about nature as is Hume's simpler subjectivism. Propensity theories of probability propose a physical property that cannot be recorded and does not necessitate or preclude any occurrence. Limiting-frequency-accounts parse probability claims as statements about the limiting relative frequency of a property in an (ill-specified) infinite sequence, but any limiting-frequency claim is consistent with any claim about any finite collection of events, and so entails nothing about what will happen at any time. Provided some way to fix their reference, limiting frequency-probability claims would at least say something about limiting properties of sequences, which is more than the alternative interpretations offer, but rather less than what we hope for from science.

2. Probabilistic Theories and Their Content: The Example of the Rasch Model

Theories whose content consists entirely of probability claims were once rare and now abound. Consider psychometrics, where until about 1960 theories of mental abilities typically assumed that there is a linear relation between how an individual performs on a psychometric test and values of various unmeasured factors, some of which also influence how that individual performs on other tests and some of which are peculiar to a single test. It was generally assumed that all factors operate in the same way for all individuals, although the values of the latent traits, and in some cases the values of the linear coefficients, may differ from individual to individual. Psychometric theories of this kind had a definite non-probabilistic content. Probability entered only in the form of assumptions about how values of the unobserved causal variables are distributed. The distributional claims--normality, for example--were essential in order to use statistical methods to make inferences about the values of linear coefficients, the number of unobserved causes, and other matters, but they were ancillary to the content of the theories. Models of this kind are still common.

In 1960 Georg Rasch published a remarkable and influential book that changed the form, and the content, of many psychometric theories. Rasch says he drew his inspiration from statistical mechanics and quantum mechanics, although his theory is more thoroughly probabilistic than either. Rasch supposed that two parameters are associated with a psychometric battery of tests and a population of people who take, or might take, the battery. One parameter varies from test item to test item and measures the difficulty of each; for each test item, that parameter has the same value for all persons. The other parameter varies from person to person and, for each person, measures the ability of that person; for each person the parameter is constant across all test items.

Rasch's framework is used in related ways to produce "models" of several kinds of tests. For a reading test that is essentially a battery in which a subject may make any number up to n errors in n items the Rasch model says that the probability a person makes k errors is given by a Poisson distribution whose single parameter is the ratio of the difficulty and ability parameters. That is the entire content of the theory. There are no underlying functional dependencies postulated, only probability relations. This, presumably, is the kind of theory philosophers of science who write of "probabilistic causality" have in mind. Although Rasch himself described orthodox estimation and test statistics for his models, the Rasch model has been given Bayesian extensions by postulating prior distributions on the Rasch parameters and letting the Rasch model describe the likelihoods.

Rasch's theory says that within the variation allowed by a psychometric battery, a person's ability is not relative to a test item and the difficulty of a test item is not relative to persons. Beyond that, the theory seems to say nothing about what happens when people take a psychometric test. If we interpret the probabilities of the Rasch models as limiting relative frequencies, the theory says nothing, literally nothing, about any observed population, no matter how large. If we interpret the probabilities of the Rasch models as subjective, then such a model is only strange and implausible autobiography. The theorist is then saying that if she knew the values of the Rasch parameters were such and such, she would have a particular distribution of degrees of belief about the errors a test subject would make. The theory says absolutely nothing about why the theorist should have those degrees of belief, about the grounds or reasons for them, nothing about the tests or the subjects or their relations that make those degrees of belief plausible, only that they are what they are. What ever kind of theory this is, it is not science.

3. Instrumental Probability

The logical oddity of Rasch's theory suggests that probability has some instrumental role, and that the real content of the theory is hidden in methods of data analysis that use probabilities. Instrumentalism about some part of language is the doctrine that sentences formulated within it are not claims but linguistic devices used to make inferences about real claims formulated in other parts of language. Skepticism about atoms, for example, was initially epistemic--facts about atoms could not be obtained from a kind of evidence relative to a vague but fixed set of assumptions. Implausible attempts were made to turn that epistemic problem into a semantic difficulty, but the transition never worked. A case has been made for instrumental mathematics, on the grounds that mathematics is a separable addition to the rest of science, which for philosophical purposes can be formulated without using mathematical terminology. The case for instrumental probability is more robust and more semantic; it rests on the fact that probability claims say nothing about what happens or could happen in the domain under scientific study, not just on the fact that they say nothing about what could be observed to happen. How, then, can probability claims be an instrument for inferences from data?

Mathematical statistics has in some measure answered this question and avoided the foundational perplexities of probability with a kind of semi-instrumentalism that uses two ingenious bridges permitting traffic only in opposite directions, one from nature to probability and the other from probability to judgements about nature. The bridge in one direction is the theory of estimation, whether in its Fisherian or Bayesian form, which provides rules connecting what happens with probability claims (either by calculating posterior distributions or through the sampling theory of estimators). The other direction is bridged by the theory of decision-making which specifies rules for action (including deciding what to believe) as a function of the subjective (or other) probabilities and the utilities of the actor. Probability and utility can be used to calculate what to say about what will happen, what would happen if ..., and so on.

Mathematical statistics and decision theory leave opaque the claims of probability in the content of science, but they clarify immensely the role of probability in the analysis of data. There is a price. The justification for the circuitous route is that, given the evidence, given the supposition that the reasoner was rational before acquiring the evidence, and given the utilities of the reasoner, the judgements that result are required by rationality. The clarification requires a change in the primary goal of inquiry form the pursuit of truth to the pursuit of rationality.

These remarks, which contain nothing novel, suggest a number of questions: The first is historical: Why and how did the goal of rationality displace the goal of informative truth in the analysis of data? I will offer what can only be a conjecture, but I hope a conjecture that, while at first surprising, becomes plausible after reflection. The second question is philosophical: Besides the detour through decision theory, is there an alternative that explains how the mathematics of probability can be used to analyze data and that keeps foremost the goal of troth rather than the goal of rationality? And the third is interpretive: Can applied statistical practice be understood as part of such an enterprise?

4. Computation and the Probabilistic Revolution in Data Analysis

We are so accustomed to treating variation in observations by statistical methods that it is hard to imagine any other way of proceeding, but it was not always so, and how it came to be is something of a puzzle. Modem methods of data analysis are usually traced to Legendre's introduction of least squares in the Appendix to his essay from 1805 on the orbits of comets. Within ten years of its appearance, the method of least squares was used in astronomy and geodesy throughout Europe. Stephen Stigler's excellent history of statistics begins with a discussion of the emergence of least-squares methods for parameter estimation and the probabilistic justification of those methods through the theory of the normal distribution. Stigler asks: "What were the characteristics of the problems faced by eighteenth-century astronomers and geodesists that led to the method's introduction and easy acceptance?" Stigler answers his own question in an historian's way, by describing the variety of other methods that emerged in the 18th century at the hands of people such as Cotes, Euler, and Mayer. Stigler's implicit answer is that in view of these precedents the emergence and acceptance of least squares is less surprising. I think a more specific answer may occur to anyone who reads Stigler's history alert to issues of computational complexity. That answer, which I will develop here, is different in emphasis from Stigler's, although not in any way incompatible with what he says. My analysis makes no pretense to original scholarship.

More than half a century before Legendre wrote, Jacques Cassini and Roger Boscovitch had used a non-probabilistic method to test hypotheses against inconsistent observations, the method now sometimes called in engineering texts "uncertain but bounded error." Their methods could in principle also have been used for parameter estimation. The idea is simple: in using a set of measurements either to test hypotheses or to estimate parameters, the investigator believes that the errors in the measurements are within some specific bound. That is, when a value x of variable X is recorded in measurement, the true value of X on that occasion is within F(x) of x, where F is some explicit function assumed known, most often, but not necessarily, a constant. The measured values of variables then determine necessary bounds on parameters to be estimated, and may exclude a hypothesis when no values of its parameters exist that are consistent with both the measurements and the error bounds assumed.

To be more definite, suppose two measurements give values Q1 and Q2, with Q1 < Q2. Then if e is the error bound for both measurements, we know the following things about the true value of Q:

$Q1 - e < Q < Q1 + e$

$Q2 - e < Q < Q2 + e$

But since Q1 < Q2, it follows that $Q2 - e < Q < Q1 + e$, which determines the interval DELTAQ of possible values of Q.

Uncertain but bounded errors propagate through calculations. Consider the "ideal gas law": pressure (P), volume (V) and temperature (T) of any sample of any gas are related by

$PV = KT$

where K is a real number that is the same for all time and for any one sample of gas, but which may differ for different samples of gases. Suppose we want to use the ideal gas law to calculate the pressure of a sample of gas at a particular time, t2. We can measure T and V at that time, but we can't use the measurements and the ideal gas law to calculate P unless we have a value for K. We measure T and V and P at some other array of values (say at time t1) for that same sample of gas and calculate K by

$K = P_1 V_1/T_1$

Since K is constant, this value of K can be used to calculate P at time t2 from measured values of V and T for that time. In other words, we calculate:

$P_2 = (P_1 V_1 T_2)/(T_1 V_2)$

Suppose the uncertainties in the other measurements are:

$V_1 = e$, $P_1 = d$, $V_2 = a$, $T_1 = g$, $T_2 = b$

Now what is the bound on the uncertainty of P_2 if we calculate P_2 from these measurements? Let's suppose all quantities are positive (so we're measuring temperature on the Kelvin scale), and to make the arithmetic easy, suppose that all measurements have numerical value 100 and all error bounds have numerical value 1. Then the most we can say is that for certain,

$95.118 < P_2 < 105.122$

Uncertainties of one part in a hundred become uncertainties of one part in ten! We can reduce the uncertainty of calculated quantities by repeated measurements of the quantities they are calculated from. Consider determining the value of K for a gas sample. We can calculate K from measurements of P, V, and T. Any one such calculation will give us error bounds on K: $K_{1min} < K < K_{1max}$. Now suppose a second set of measurements of P, V, and T are taken. Since the measurements are not perfectly accurate, PV/T will generally not be exactly equal for the two sets of measurements. Suppose, for example, that $P_2V_2/T_2 < P_1V_1/T_1$. Then if the error bounds are constant, $K_{2min} < K_{1min}$ and $K_{2max} < K_{1max}$. If $_{K2max} < K_{1min}$ the ideal gas law is refuted. Otherwise, $K_{1min} < K < K_{2max}$ is a narrower bound on K than either single set of measurements provides. In fact it is easy to see that repeated measurements can give arbitrarily accurate estimates of computed quantities provided (1) the law used to do the computation is true, and (2) the error bounds of each individual measurement are the best possible--that is, errors as large as those allowed do occur, but no larger errors occur.

The method illustrated above is similar to methods that antedate least squares. In 1755, Father Boscovitch analyzed five data points on the length of meridian arc at various latitudes, taken for the purpose of testing the Newtonian hypothesis of an ellipsoidal Earth. Boscovitch in effect had a linear equation (see Stigler, p. 42) in two unknowns and five data points, allowing 10 determinations of the unknown parameters. He in fact computed all 10 values, and also computed the average value (1/198) of one of the parameters, the ellipticity, and argued that the difference between the individual values of the ellipticity (ranging from 1/78 to -1.486) and the average was too large to be due to measurement error. He concluded that the elliptical hypothesis must be rejected.

Boscovitch's argument is valid if we accept his bounds on the errors of measurement. Stigler claims that Boscovitch's error bounds were unreasonable at the time, and using a more reasonable error bound (100 toises) he finds that the least squares line is within the error interval for all five observations. But that is only to deny Boscovitch's premise, not the validity of Boscovitch's argument. Had Boscovitch used the larger error value he could have used his procedure to estimate an interval of values for the ellipticity of the figure of the Earth.

Boscovitch's procedure is sensible, simple in description, informative about the truth or falsity of hypotheses of interest, and requires only an elementary kind of prior belief that could readily be elicited form scientific practitioners. It corresponds to an interval estimation procedure that is, if the assumptions about error bounds are correct, logically guaranteed to give an interval estimate containing the true value if one exists. Moreover, in most realistic cases, the interval estimates obtained from a number of measures would be very informative, narrowly circumscribing the true value. Boscovitch himself later abandoned the method for a procedure that minimized the sum of absolute values of differences between estimate and observations. And, after the turn of the century, uncertain but bounded error methods such as Boscovitch's were rapidly displaced by least squares, a method that had no guarantee of truth at all. Why?

Part of the reason is, of course, that within fifteen years of the appearance of Legendre's book, Gauss and Laplace gave least squares what remains the standard justification for its use: the expected value of least-squares estimates is the true value for normally distributed variables, and minimizes the expected squared error of the estimate--in modern terms least squares is the minimum-variance-unbiased estimator. I don't know just when it was realized that least squares is also the maximum-likelihood estimator for normal distributions, but maximum-likelihood ideas had already been introduced, after a fashion, by Daniel Bernoulli. The central-limit theorems justified the normal distribution: normal distributions are the limits of binomial distributions, or more substantively, the normal distribution results in the limit from summing appropriately small, unrelated causes.

Despite these arguments, some further motive seems required to explain the rapid adoption of least squares, for despite its elegance and mathematical pedigree, this justification for least squares sidestepped the question of how most reliably to infer the true value of an unmeasured parameter from inconsistent premises consisting of an equation containing the parameter and multiple observed values of other quantities in the equation. Then as now, probability had no evident connection with truth, and while least squares had an evident connection with minimizing expected squared loss, it had a less evident connection with estimating the truth. Probability could not be reduced to any natural property that could be determined from observation; despite Bayes's and Laplace's efforts, no logical connections had been established between probability as a measure of credence and truth; and the limiting connections between truth, probability and frequency were not only in the limit, they were circular. Bemoulli's theorem, for example, did not say that a binomial probability could be determined in the limit from a sequence of trials; the theorem was twice circular, requiring that the trials be independent and giving convergence only in probability. In contrast, under assumptions investigators were disposed to make, the method of uncertain but bounded error could be applied directly to measurements rather than to distributions, and elements of that application were used--especially to test hypotheses--until the appearance of least squares.

I suggest that, besides having a justification that called forth the great mathematics and mathematicians of the age, and besides giving intuitive results, least squares had an essential virtue that uncertain but bounded error did not: least squares is computationally tractable. The complexity of least-squares computations is linear in the number of data points and bounded above by the cube of the number of unknowns. With k unknowns, Boscovitch's procedure requires solving all possible sets of k equations that can be formed from n data points, or n choose k cases. So the number of computations required by Boscovitch's method is on the order of 2k to the n choose k power. There may, of course, be more efficient methods of estimation or of testing by uncertain but bounded error than Boscovitch's, methods that reduce the time complexity to a polynomial in the number of observations. If so, the methods are not obvious. I guess, to the contrary, that uncertain but bounded estimation of (rational) values of k coefficients in a linear equation with $r > k$ variables from $n > k$ observations of (rational) values is NP hard (in n) for all $k > 2$, although I have no reduction to offer. (Even if my guess is true for the most informative uncertain but bounded estimate, it is obviously not true for estimates that are less informative--that produce wider estimates of the possible values of the parameters of interest. Indeed any set of k measurements will produce bounds on the parameter estimates that include the best bounds produced by a more informative procedure.)

In general, there is little to be said in favor of arbitrarily throwing away information, but in particular cases it may be evident that certain sets of measured values will produce the tightest estimated bounds on specific parameters, while other estimated values will produce the tightest bounds on other parameters, and so on. Where this is so, heuristic selection of the subsets to be used for estimating specific parameters may produce estimated sets of values for parameters that are reasonably close to those that would be produced by a full procedure such as Boscovitch's, and, indeed, may be small enough that a point estimate reasonably conveys the information about estimated values of parameters. Some of the proposals Stigler describes that intermediate between uncertain but bounded error and least squares involve just such heuristic selection of subsets of measurements, or other devices that reduce the complexity of the calculation by neglecting information. Avoiding computational complexity appears to have been the principal motivation for these innovations. Thus Euler, for example, tried to estimate eight free parameters of unknown but constant values in a linear equation with 7 measured variables for which 75 observations were available. Although from various passages it seems likely that Euler thought of the problem in terms of uncertain but bounded error, for complexity reasons he obviously could not follow Boscovitch's example. Instead he determined values for two of the eight unknown parameters by noting that the coefficients of the other six parameters were periodic with a period of 59 years. Subtracting values for two pairs of observations, one element of each pair taken a multiple of 59 years from the other observation of the pair, he obtained two equations in the two unknown parameters, which he solved. The result was not sensitive to the pairs chosen. That left six unknowns. Euler tried the same strategy for the other unknowns but found inconsistent values because there were no periodicities that exactly separated the other parameters. Stigler reports that Euler essentially gave up on the problem. Euler's chief idea appears to have been to combine estimation with induction: by finding additional independent constraints in the data, values for parameters can be determined and the complexity of the estimation problem exponentially reduced. That is a good but perhaps not often applicable idea.

A year after Euler's work appeared, in addressing a problem from lunar theory, Johann Mayer obtained a linear equation in three unknown parameters--I will refer to them as a, b, and c--and three measured variables for which he had 27 observations. Again, Boscovitch's method is infeasible. Instead, substituting each observation into the equation, Mayer formed 27 equations in 3 unknowns. He divided the 27 into three groups of nine equations each, putting in one group the equations with the largest positive values for the coefficient of b, putting in a second group the equations with the nine largest negative values for the coefficient of b, and putting the remaining equations in a third group, which Mayer claimed (incorrectly) had the largest values for the coefficient of c. Mayer then summed the equations in the respective groups to obtain three equations in the three unknowns, which he then solved.

Replacing a set of equations in the same variables by the equation of their sums evidently decreases the complexity of the problem exponentially, but it can and typically will lead to errors in the estimate, and from the perspective of uncertain but bounded error one would like to know the bounds on those errors. Mayer himself seems to have endorsed this perspective, and, giving error bounds on individual observations, proposed that bounds on the errors of his estimates could be calculated by following them through his procedure, although he did not give the calculation, which would require a good deal of work.

The structure of complexity for estimation and for testing need not be the same. In testing a linear hypothesis, for example, we need only consider the family of lines passing within the error bounds of the most extreme observations, and determine whether at least one of those lines intersects the error region around every other observation. If so, the hypothesis is consistent with the observations, and otherwise not. The procedure is polynomial in the number of observations, and generalizes to more dimensions. Boscovitch's procedure, which may be necessary for estimation, is unnecessarily complex for hypothesis testing. That had in effect been discovered by Jacques Cassini prior to Boscovitch's work.

In 1740, Cassini discussed 16 observations of the obliquity of the ecliptic over a span of nearly 2,000 years. Cassini was interested in the hypotheses that the obliquity changes linearly with time. He analyzed the data by fitting a straight line to the first and last observations, and then computing the interpolated values for each recorded time of measurement. Cassini claimed that the observed and interpolated values were too great to be due to measurement error, objecting especially to the difference between interpolated values and those given by Ptolemy and Pappus. Goldstein has shown that, to the contrary, Ptolemy's value was influenced by those of his predecessors and was in error by about 10'. That aside, Cassini's argument appears to have an interesting structure. He believed that none of the observations could be in error by more than a specific interval e, and he showed that some of the observations are more than that interval away from the unique line determined by the two observations that are most extreme in their values for both variables. He concluded that no line can account for the observations.

The defect I have conjectured our predecessors found in uncertain but bounded error is computational. Of course I do not claim that anyone in the 18th or 19th centuries did an explicit complexity analysis. No one had to do so in order to recognize that uncertain but bounded error estimation could not be carried out in practice, and least squares could be. In judging the plausibility of an historical explanation that places so much weight on computational considerations, it might usefully be kept in mind that in more recent times computational considerations have had a strong influence on the forms of probabilistic data analysis. Guilford's Psychometric Methods examines both Thurstone's factor analysis and a competing procedure due to Spearman. Thurstone provided an algorithm for generating linear theories that used unobserved common causes to account for rank constraints exhibited in the covariances among measured variables. Spearman had a technique for theory generation based on accounting for rank-one constraints. Guilford recommended Thurstone's method over Spearman's on the grounds that Spearman's procedures were computationally intractable. More recently, with the computer and computer algorithms, the argument has been reversed, yielding claims that iterative fitting of covariances models is so computationally demanding that it limits search, and adaptations of Spearman's methods are actually more feasible. Again, Bayesian methods have been slow to be adapted partly for philosophical reasons, but perhaps more importantly because the circumstances in which posterior distributions can be computed from prior distributions and data have been very limited. Very recently there has been a lively interest in an adaptation of ergodic theorems to produce posterior distributions by simulation, but in practice these "Gibb's sampler" procedures so far remain too computationally demanding to be very practical.

Before we leave uncertain but bounded error, there is a last, irresistible confusion to consider: that assertions of uncertain but bounded error are limiting cases of assertions of a probability that a quantity lies in some specified set of values. The idea is that simple uncertain but bounded error claims are really claims of probabilistic certainty, probability 1 claims. The more general form, according to this idea, is "the probability is p that the value of X lies within epsilon of x." The idea is wrong. The assertion "X lies within epsilon of x" does not entail, nor is it entailed by, "the limiting relative frequency of cases in which X lies within epsilon of x is one," nor by "my degree of belief is one that X lies within epsilon of x." The sentence "X lies within epsilon of x" has no logical connection with any interesting probability claim.

5. Approximation, Distributional Forms and Finite Frequencies

In contrast to the philosophical accounts of probability, there is an unphilosophical account often advanced by practitioners and teachers that does connect probability claims directly with claims about what happens, and that connection may explain the popularity of the proposal. The "long-run frequency" interpretation of probability, popular in textbooks, says the probability of an outcome in a trial is approximately the relative frequency of the outcome in a "long" sequence of trials. A non-sequential version of the same idea is that probability claims are about the approximate relative frequency distribution of values of features of individuals in large populations (of whatever). For apparently good reasons, the finite-frequency account is short on philosophical champions: it is twice vague, since it does not specify the meaning of "approximate" or of "long"; for any way of making the finite-frequency claim precise, typical literal probability claims (e.g., the claim that a distribution is normal) are false; and the finite-frequency interpretation seems committed to the Gambler's Fallacy, for if "the probability of heads is 1/2" means something like "in a sequence of 1000 tosses or more the frequency of heads will be nearly 1/2" then the gambler who find that a coin he believes to be fair has turned up heads in 500 consecutive tosses can reasonably infer that tails are due: unless almost all of the remaining tosses are tails the probability claim, under the finite-frequency interpretation, will be false.

I recommend charity to gamblers and to practitioners. I suggest that the finite-frequency story is something else besides a definition of "probability,'" that it is a compressed account of how inferences from data may be made with the aid of the mathematics of probability, but without the obscure thing itself. I think we are more faithful to practice if we understand the finite-frequency interpretation as a proposal to use the language and mathematics of probability to approximately describe actual or potential finite populations, and as a means of generating definite, non-probabilistic hypotheses. Thus, to illustrate the first point, the claim that "adult human male height has a normal probability distribution" is a way of saying that a histogram of the relative frequencies of such heights would be approximated by the curve of the normal density function. The claim may be compared to "this sample of gas approximately satisfies the ideal gas law." Both have an unspecified parameter--the degree of approximation--and both are properties of a collective, the entire population in the case of probability and pairs of P, V, T states in the case of the gas law. To the objection that informed people do not behave as though they were committing the distributional analogue of the gambler's fallacy, I say that so far as that is true it is so because in such cases informed people reject the distributional assumption. When the sample is large but the distribution wants a tail, we reject the proposition that in a still larger sample the distribution will be normal. And when we are not allowed to reject the distributional assumption and we are dealing with inferences about a finite population from samples taken without replacement from that population, the gambler's fallacy is no fallacy at all, only good logic.

One might reasonably wonder whether there is any gain in content in moving from claims about probability to claims about the "approximate" form of distribution of an unspecifiedly large sample. Unlike the probability claim itself, the degree of approximation in an approximate empirical frequency claim can be made explicit and empirical in various ways. Analytic measures of approximation can be used to generate claims that are perfectly definite and perfectly empirical, even though they may be clothed in the language of probability. For example, one may say that the sum of squared vertical distances from bar heights on a frequency histogram and points on some normal distribution is less than some specific number, or that for each point in the histogram the distance to the normal curve is less than a percentage of the height of histogram, and so on. Again, uncertain but bounded error in measurements will create uncertain but bounded errors in the points on the histogram, and thus a family of normals that might, within any measure of approximation, describe the histogram. When made explicit in these or other ways, "probability" claims become assertions about bounds on arrangements of values of quantities in finite populations. They become an especially interesting variety of claims, not about probabilities, but about uncertain but bounded errors in finite frequency distributions. Entirely explicit versions of finite frequency claims of probability, on my analysis, are claims about the uncertain but bounded error of some function of the empirical distribution of a quantity (or quantities) in an actual or potential finite population.

The Rasch models and other theories whose claims appear to be probabilistic have some content if we understand the probabilities involved to be hypotheses about the approximate distributions of values in large finite samples. That interpretation may make the content of theories vague, but vagueness is better than vacuity. The interpretation may violate the mathematics of probability, but that is essential to the very idea of probability theory as an instrument.

One scientific value of approximate distributional hypotheses lies in the fact that variables in systems that satisfy approximately no informative restriction on n arbitrary observed values, for small n, may nonetheless to useful approximations satisfy distributional hypothesis for large n. The claim that a finite population of values is distributed approximately as some probability density sometimes conveys important extra information that could not be given by saying that some quantity or quantities have specific values, approximately. Independence hypotheses provide one important example, fundamental to experimental design, where approximate distributions can be much more informative about parameters of interest than can uncertain but bounded error claim about the values of quantities whose distribution is described.

Consider an experiment to test a simple linear hypothesis:

$Y = aX + e$

The object is to estimate the parameter a. We will suppose that values of X are determined deliberately by some mechanism that is believed to distribute X approximately independently of e. Suppose we assume that X and Y are measured perfectly. With uncertain but bounded error nothing can be inferred about the value of parameter a because e is unknown. We can only estimate a if we assume some definite, finite set of possible values for e without the help of any measurement of that quantity.

The treatment with approximate distributions is quite different. Because X is a treatment variable under experimental control, all experimental design assumes that e and X are (approximately) independently distributed. Then a can be estimated in various ways, for example from estimates of the variances and covariances related to the parameter a by the familiar derivation:

$\mathrm{Exp}(X,Y) = a\ \mathrm{Exp}(X,X) + \mathrm{Exp}(X,e) = a\ \mathrm{EXP}(X,X) + \mathrm{COV}(X,e) + \mathrm{Exp}(X)\mathrm{Exp}(e) = a\ \mathrm{Exp}(X,X) + \mathrm{Exp}(X)\mathrm{Exp}(e) = a\ \mathrm{Exp}(X,X) + \mathrm{Exp}(X)\mathrm{Exp}(Y) - a\mathrm{Exp}^2(X)$ (Because $\mathrm{Exp}(X)\mathrm{Exp}(Y) = a\mathrm{Exp}^2(X) + \mathrm{Exp}(X)\mathrm{Exp}(e)$) $\mathrm{Exp}(X,Y) - \mathrm{Exp}(X)\mathrm{Exp}(Y) = a\ (\mathrm{Exp}(X,X) - \mathrm{Exp}^2(X))$ $\mathrm{COV}(X,Y) = a\ \mathrm{Var}(X)$

Once one thinks of approximate descriptions of real or potential finite frequency distributions, the relation between the absence of causal connection and approximate independence or conditional independence opens a huge space of applications unavailable without distributions.

6. Concluding Remarks

I have tried to say how the theory of probability, which, much as Locke's substance, is about something we know not what, has provided and can provide instruments for making valid inferences about the values of quantities and the truth or falsity of hypotheses, rather than valid inferences as to which decision has the highest expected utility. The sometimes bitter debates between those who describe themselves as frequentists and those who describe themselves as subjective Bayesians has often turned on charges by the former that the latter abandon the "objectivity" of science and by the latter that the former dissemble about the "subjectivity" of their probability judgements. My belief is that, among statisticians anyway, the dispute often confuses content with justification. The "objectivity" of the frequentists is in the content of their probability judgements, which, while usually stated as about an un-empirical probability, are often really vague empirical claims about finite frequencies. That sort of objectivity is genuinely lost in subjective Bayesian interpretations. The "subjectivity" kept hidden by frequentists is that there is often no explicit justification beyond their own opinion for aspects of their empirical claims. That subjectivity can be made entirely explicit without sacrificing the objective--that is empirical--content of frequency claims, and its recognition does not require, or even invite, recourse to subjective probability. Bayesian criticisms do address a confused and uncertain frequentist statistical practice, in which the point of making empirical claims is often forgotten or fudged. Data analysis might look somewhat different, and from a logical perspective, better, if "orthodox" or "frequentist" or "objectivist" statisticians were candid and consistent about the instrumental use they would make of probability, and if probability were understood as a tool for idealization and approximation rather than a normative scold.

REFERENCES

Guilford, J. (1936) Psychometric Methods. New York: McGraw-Hill.

Rasch, G. (1960) "Probabilistic Models for Some Intelligence and Attainment Tests". Copenhagen: Studies in Mathematical Psychology.

Stigler, S. (1986) The History of Statistics. Cambridge, MA: Harvard University Press

~~~~~~~~

By Clark Glymour

Carnegie Mellon University, Institute for Human and Machine Cognition, University of West Florida, University of California, San Diego

Source: Monist, Apr2001, Vol. 84 Issue 2, p284, 17p
Item: 4562569
~~~~~~~~