Tour IV More Auditing: Objectivity and Model Checking

4.8 All Models Are False

> . . . it does not seem helpful just to say that all models are wrong. The very word model implies simplification and idealization. . . . The construction of idealized representations that capture important stable aspects of such systems is, however, a vital part of general scientific analysis. (Cox 1995, p. 456)

A popular slogan in statistics and elsewhere is "all models are false!" Is this true? What can it mean to attribute a truth value to a model? Clearly what is meant involves some assertion or hypothesis about the model – that it correctly or incorrectly represents some phenomenon in some respect or to some degree. Such assertions clearly can be true. As Cox observes, "the very word model implies simplification and idealization." To declare, "all models are false" by dint of their being idealizations or approximations, is to stick us with one of those "all flesh is grass" trivializations (Section 4.1). So understood, it follows that all statistical models are false, but we have learned nothing about how statistical models may be used to infer true claims about problems of interest. Since the severe tester's goal in using approximate statistical models is largely to learn where they break down, their strict falsity is a given. Yet it does make her wonder why anyone would want to place a probability assignment on their truth, unless it was 0? Today's tour continues our journey into solving the problem of induction (Section 2.7).

Assigning a probability to either a substantive or a statistical model is very different from asserting it is approximately correct or adequate for solving a problem. The philosopher of science Peter Achinstein had hoped to discover that his scientific heroes, Isaac Newton and John Stuart Mill, were Bayesian probabilists, but he was disappointed; what he finds is enlightening:

> Neither in their abstract formulations of inductive generalizations (Newton's rule 3; Mill's definition of 'induction') nor in their examples of particular inductions to general conclusions of the form 'all As are Bs' does the term 'probability' occur. Both write that from certain specific facts we can conclude general ones – not that we can conclude general propositions with probability, or that general propositions have a probability . . . From the inductive premises we simply conclude that the generalization is true, or as Newton allows in rule 4, 'very nearly true,' by which he appears to mean not 'probably

> true' but 'approximately true' (as he does when he takes the orbits of the satellites of Jupiter to be circles rather than ellipses). (Achinstein 2010, p. 176)

There are two main ways the "all models are false" charge comes about:

1. The statistical inference refers to an idealized and partial representation of a theory or process.
2. The probability model, to which a statistical inference refers, is at most an idealized and partial representation of the actual data-generating source.

Neither of these facts precludes the use of these *false* models to find out true things, or to correctly solve problems. On the contrary, it would be impossible to learn about the world if we did not deliberately falsify and simplify.

Adequacy for a Problem. The statistician George Box, to whom the slogan "all models are wrong" is often attributed, goes on to add "But some are useful" (1979, p. 202). I'll go further still: all models are false, no useful models are true. Were a model so complex as to represent every detail of data "realistically," it wouldn't be useful for finding things out. Let's say a statistical model is useful by being adequate for a problem, meaning it may be used to find true or approximately true solutions. Statistical hypotheses may be seen as conjectured solutions to a problem. A statistical model is adequate for a problem of statistical inference (which is only a subset of uses of statistical models) if it enables controlling and assessing if purported solutions are well or poorly probed, and to what degree. Through approximate models, we learn about the "important stable aspects" or systematic patterns when we are in the context of phenomena that exhibit statistical variability. When I speak of ruling out mistaken interpretations of data, I include mistakes about theoretical and causal claims. If you're an anti-realist about science, you will interpret, or rather reinterpret, theoretical claims in terms of observable claims of some sort. One such anti-realist view we've seen is instrumentalism: unobservables including genes, particles, light bending may be regarded as at most instruments for finding out about observable regularities and predictions. Fortunately we won't have to engage the thorny problem of realism in science, we can remain agnostic. Neither my arguments, nor the error statistical philosophy in general, turn on whether one adopts one of the philosophies of realism or anti-realism. Today's versions of realism and anti-realism are quite frankly too hard to tell apart to be of importance to our goals. The most important thing is that both realists and non-realists require an account of statistical inference. Moreover, whatever one's view of scientific theories, a statistical analysis of problems of actual experiments involves abstraction and creative analogy.

Testing Assumptions is Crucial. You might hear it charged that frequentist methods presuppose the assumptions of their statistical models, which is puzzling because when it comes to testing assumptions it's to frequentist methods that researchers turn.

> It is crucial that any account of statistical inference provides a conceptual framework for this process of model criticism, . . . the ability of the frequentist paradigm to offer a battery of simple significance tests for model checking and possible improvement is an important part of its ability to supply objective tools for learning. (Cox and Mayo 2010, p. 285)

Brad Efron is right to say the frequentist is the pessimist, who worries that "if anything can go wrong it will," while the Bayesian optimistically assumes if anything can go right it will (Efron 1998, p. 99). The frequentist error statistician is a worrywart, resigned to hoping things are half as good as intended. This also makes her an activist, deliberately reining in some portion of a problem so that it's sufficiently like one she knows how to check. Within these designated model checks, assumptions under test are intended to arise only as i-assumptions. They're assumptions for drawing out consequences, for possible falsification.

"In principle, the information in the data is split into two parts, one to assess the unknown parameters of interest and the other for model criticism" (Cox 2006a, p. 198). The number of successes in n Bernoulli trials, recall, is a *sufficient* statistic, and has a Binomial sampling distribution determined by θ, the probability of success on each trial (Section 3.3). If the model is appropriate then any permutation of the r successes in n trials has a known probability. Because this conditional distribution (X given s) is known, it serves to assess if the model is violated. If it shows statistical discordance, the model is disconfirmed or falsified. The key is to look at residuals: the difference between each observed value and what is expected under the model. (We illustrate with the runs test in Section 4.11.) It is also characteristic of error statistical methods to be relatively robust to violation.

Central Limit Theorem. In the presentation on justifying induction (Section 2.7), we heard Neyman stress how the empirical Law of Large Numbers (LLN) is in sync with the mathematical law in a number of "real random experiments." Supplementing the LLN is the Central Limit Theorem (CLT). It tells us that the mean $\overline{X}$ of n independent random variables, each X with mean μ, and finite non-zero σ^2, is approximately Normally distributed with its mean equal to μ and standard deviation $\sigma/\sqrt{n}$ – regardless of the underlying distribution of X. So long as n is reasonably large (say 40 or 50), and the underlying distribution is not too asymmetrical, the Normal distribution gives a good approximation, and is robust for many cases where IID is violated. The CLT tells us that $\overline{X}$ standardized is N(0,1). The finite non-zero variance isn't much of a restriction, and even this has been capable of being relaxed.

The CLT links a claim or question about a statistical hypothesis to claims about the relative frequencies that would be expected in applications (real or hypothetical) of the experiment. Owing to this link, we can use the sample mean to inquire about values of μ that are capable or incapable of bringing it about. Our standardized difference measures observed disagreement, and classifies those improbably far from hypothesized values. Thus, statistical models may be adequate for real random experiments, and hypotheses to this effect may pass with severity.

Exhibit (xii): Pest Control. Neyman (1952) immediately turns from the canonical examples of real random experiments – of coin tossing and roulette wheels – to illustrate how "the abstract theory of probability ... may be, and actually is, applied to solve problems of practice importance" such as pest control (p. 27)! Given the lack of human control here, he expects the mechanism to be complicated. The first attempt to model the variation in larvae hatched from moth eggs is way off.

> [I]f we attempt to treat the distribution of larvae from the point of view of [the Poisson distribution], we would have to assume that each larva is placed on the field independently of the others. This basic assumption was flatly contradicted by the life of larvae as described by Dr. Beall. Larvae develop from eggs laid by moths. It is plausible to assume that, when a moth feels like laying eggs, it does not make any special choice between sections of a field planted with the same crop and reasonably uniform in other respects. (1952, pp. 34–5).

So it's plausible to suppose the Poisson distribution for the spots where moths lay their eggs. However, a data analysis made it "clear that a very serious divergence exists" between the actual distribution of larvae and the Poisson model (ibid., p. 34). Larvae expert, Dr. Beall, explains why: At each "sitting" a moth lays a whole batch of eggs and the number of eggs varies from one cluster to another. "After hatching ... the larvae begin to look for food and crawl around" but given their slow movement "if one larva is found, then it is likely that the plot will contain more than one from the same cluster" (ibid., p. 35). An independence assumption fails. (I omit many details; see Neyman 1952, Gillies 2001.)

The main thing is this: The misfit with the Poisson model leads Neyman to arrive at a completely novel distribution: he called it the type A distribution (a "contagious" distribution). Yet Neyman knows full well that even the type A distribution is strictly inadequate, and a far more complex distribution would be required for answering certain questions. He knows it's strictly false. Yet it suffices to show why the first attempt failed, and it's adequate to solving his immediate problem in pest control.

Souvenir U: Severity in Terms of Problem-Solving

The aim of inquiry is finding things out. To find things out we need to solve problems that arise due to limited, partial, noisy, and error-prone information. Statistical models are at best approximations of aspects of the data-generating process. Reasserting this fact is not informative about the case at hand. These models work because they need only capture rather coarse properties of the phenomena: the error probabilities of the test method are approximately and conservatively related to actual ones. A problem beset by variability is turned into one where the variability is known at least approximately. Far from wanting true (or even "truer") models, we need models whose deliberate falsity enables finding things out.

Our threadbare array of models and questions is just a starter home to grow the nooks and crannies between data and what you want to know (Souvenir E, Figure 2.1). In learning about the large-scale theories of sexy science, intermediate statistical models house two "would-be" claims. Let me explain. The theory of GTR does not directly say anything about an experiment we could perform. Splitting off some partial question, say about the deflection effect, we get a prediction about what *would be* expected were the deflection effect approximately equal to the Einstein value, 1.75". Raw data from actual experiments, cleaned and massaged, afford inferences about intermediate (astrometric) models; inferences as to what it would be like were we taking measurements at the limb of the sun. The two counterfactual inferences – from the theory down, and the data up – meet in the intermediate statistical models. We don't seek a probabilist assignment to a hypothesis or model. We want to know what the data say about a conjectured solution to a problem: What erroneous interpretations have been well ruled out? Which have not even been probed? The warrant for these claims is afforded by the method's capabilities to have informed us of mistaken interpretations. *Statistical methods are useful for testing solutions to problems when this capability/incapability is captured by the relative frequency with which the method avoids misinterpretations.*

If you want to avoid speaking of "truth" you can put the severity requirement in terms of solving a problem. A claim *H* asserts a proposed solution S to an inferential problem is adequate in some respects. It could be a model for prediction, or anything besides.

H: S is adequate for a problem

To reject *H* means "infer *S* is inadequate for a problem." If none of the possible outcomes lead to reject *H* even if *H* is false – the test is incapable of finding inadequacies in *S* – then "do not reject *H*" is BENT evidence that *H* is true. We move from no capability, to some, to high:

> If the test procedure (which generally alludes to a cluster of tests) very rarely rejects *H*, if *H* is true, then "reject *H*" provides evidence for falsifying *H* in the respect indicated.

You could say, a particular inadequacy is corroborated. It's still an inferential question: what's warranted to infer. We start, not with hypotheses, but questions and problems. We want to appraise hypothesized answers severely.

I'll meet you in the ship's library for a reenactment of George Box (1983) issuing "An Apology for Ecumenism in Statistics."

4.9 For Model-Checking, They Come Back to Significance Tests

> Why can't all criticism be done using Bayes posterior analysis . . .? The difficulty with this approach is that by supposing all possible sets of assumptions known *a priori*, it discredits the possibility of new discovery. But new discovery is, after all, the most important object of the scientific process. (George Box 1983, p. 73)

Why the apology for ecumenism? Unlike most Bayesians, Box does not view induction as probabilism in the form of probabilistic updating (posterior probabilism), or any form of probabilism. Rather, it requires critically testing whether a model M_i is "consonant" with data, and this, he argues, demands frequentist significance testing. Our ability "to find patterns in discrepancies $M_i - y_{\mathrm{d}}$ between the data and what might be expected if some tentative model were true is of great importance in the search for explanations of data and of discrepant events" (Box 1983, p. 57). But the dangers of apophenia raise their head.

> However, some check is needed on [the brain's] pattern seeking ability, for common experience shows that some pattern or other can be seen in almost any set of data or facts. This is the object of diagnostic checks and tests of fit which, I will argue, require frequentist theory significance tests for their formal justification. (ibid.)

Once you have inductively arrived at an appropriate model, the move, on his view, "is entirely *deductive* and will be called *estimation*" (ibid., p. 56). The