Souvenir D: Why We Are So New

What's Old? You will hear critics say that the reason to overturn frequentist, sampling theory methods – all of which fall under our error statistical umbrella – is that, well, they've been around a long, long time. First, they are scarcely stuck in a time warp. They have developed with, and have often been the source of, the latest in modeling, resampling, simulation, Big Data, and machine learning techniques. Second, all the methods have roots in long-ago ideas. Do you know what is really up-to-the-minute in this time of massive, computer algorithmic methods and "trust me" science? A new vigilance about retaining hard-won error control techniques. Some thought that, with enough data, experimental design

[6] From Cox and Hinkley 1974, p. 51. The likelihood function corresponds to the normal distribution of X around μ with SE $\sigma/\sqrt{n}$. The likelihood at $\mu = 0$ is $\exp(-0.5k^2)$ times that at $\mu = \bar{x}$. One can choose k to make the ratio small. "That is, even if in fact $\mu = 0$, there always appears to be strong evidence against $\mu = 0$, at least if we allow comparison of the likelihood at $\mu = 0$ against any value of μ and hence in particular against the value of μ giving maximum likelihood". However, if we confine ourselves to comparing the likelihood at $\mu = 0$ with that at some fixed $\mu = \mu'$, this difficulty does not arise.

could be ignored, so we have a decade of wasted microarray experiments. To view outcomes other than what you observed as irrelevant to what $\boldsymbol{x}_0$ says is also at odds with cures for irreproducible results. When it comes to cutting-edge fraud-busting, the ancient techniques (e.g., of Fisher) are called in, refurbished with simulation.

What's really old and past its prime is the idea of a logic of inductive inference. Yet core discussions of statistical foundations today revolve around a small cluster of (very old) arguments based on that vision. Tour II took us to the crux of those arguments. Logics of induction focus on the relationships between given data and hypotheses – so outcomes other than the one observed drop out. This is captured in the Likelihood Principle (LP). According to the LP, trying and trying again makes no difference to the probabilist: it is what someone intended to do, locked up in their heads.

> It is interesting that frequentist analyses often need to be adjusted to account for these 'looks at the data,'... That Bayesian analysis claims no need to adjust for this 'look elsewhere' effect – called the *stopping rule principle* – has long been a controversial and difficult issue... (J. Berger 2008, p. 15)

The irrelevance of optional stopping is an asset for holders of the LP. For the task of criticizing and debunking, this puts us in a straightjacket. The warring sides talk past each other. We need a new perspective on the role of probability in statistical inference that will illuminate, and let us get beyond, this battle.

New Role of Probability for Assessing What's Learned. A passage to locate our approach within current thinking is from Reid and Cox (2015):

> Statistical theory continues to focus on the interplay between the roles of probability as representing physical haphazard variability ... and as encapsulating in some way, directly or indirectly, aspects of the uncertainty of knowledge, often referred to as epistemic. (p. 294)

> We may avoid the need for a different version of probability by appeal to a notion of calibration, as measured by the behavior of a procedure under hypothetical repetition. That is, we study assessing uncertainty, as with other measuring devices, by assessing the performance of proposed methods under hypothetical repetition. Within this scheme of repetition, probability is defined as a hypothetical frequency. (p. 295)

This is an ingenious idea. Our meta-level appraisal of methods proceeds this way too, but with one important difference. A key question for us is the proper epistemic role for probability. It is standardly taken as providing a probabilism, as an assignment of degree of actual or rational belief in a claim, absolute or comparative. We reject this. We proffer an alternative theory: a severity assessment. An account of what is warranted and unwarranted to infer – a normative epistemology – is not a matter of using probability to assign rational beliefs, but to control and assess how well probed claims are.

If we keep the presumption that the epistemic role of probability is a degree of belief of some sort, then we can "avoid the need for a different version of probability" by supposing that good/poor performance of a method warrants high/low belief in the method's output. Clearly, poor performance is a problem, but I say a more nuanced construal is called for. The idea that partial or imperfect knowledge is all about degrees of belief is handed down by philosophers. Let's be philosophical enough to challenge it.

New Name? An error statistician assesses inference by means of the error probabilities of the method by which the inference is reached. As these stem from the sampling distribution, the conglomeration of such methods is often called "sampling theory." However, sampling theory, like classical statistics, Fisherian, Neyman–Pearsonian, or frequentism are too much associated with hardline or mish-mashed views. Our job is to clarify them, but in a new way. Where it's apt for taking up discussions, we'll use "frequentist" interchangeably with "error statistician." However, frequentist error statisticians tend to embrace the long-run performance role of probability that I find too restrictive for science. In an attempt to remedy this, Birnbaum put forward the "confidence concept" (Conf), which he called the "one rock in a shifting scene" in statistical thinking and practice. This "one rock," he says, takes from the Neyman–Pearson (N-P) approach "techniques for systematically appraising and bounding the probabilities (under respective hypotheses) of seriously misleading interpretations of data" (Birnbaum 1970, p.1033). Extending his notion to a composite alternative:

> Conf: An adequate concept of statistical evidence should find strong evidence against H_0 (for $\sim H_0$) with small probability α when H_0 is true, and with much larger probability $(1 - \beta)$ when H_0 is false, increasing as discrepancies from H_0 increase.

This is an entirely right-headed pre-data performance requirement, but I agree with Birnbaum that it requires a reinterpretation for evidence post-data (Birnbaum 1977). Despite hints and examples, no such evidential interpretation has been given. The switch that I'm hinting at as to what's required for an evidential or epistemological assessment is key. Whether one uses a frequentist or a propensity interpretation of error probabilities (as Birnbaum did) is not essential. *What we want is an error statistical approach that controls and assesses a test's stringency or severity.* That's not much of a label. For short, we call someone who embraces such an approach a severe tester. For now I will just venture that a severity scrutiny illuminates all statistical approaches currently on offer.