

Liam Kelly
STS 6334
Final Paper
12/12/02

Incommensurable Or Merely Incompatible?:
Bayesianism and Error Statistics Reconsidered

Within the history and philosophy of statistics, perhaps no conflict has been as deeply entrenched as the debate between the Bayesian and the error-statistical approaches to scientific inference. This bad blood extends well beyond quibbles over interpretation of data. Each viewpoint seems to cast valid inference as an entirely different sort of thing, with entirely different requirements for justification. Thomas Kuhn would likely view the rift as a paradigmatic one – a rift of fundamental conceptions about the way that science ought to be performed, and one which no amount of negotiation can reconcile. While I don't accept Kuhn's paradigmatic account literally, I do think that it can be a useful tool for uncovering the true sources of conflict in a seemingly stalled scientific debate. In order to move beyond that point, however, we need to be willing to abandon a strict adherence to the paradigmatic model of scientific change, and to work harder at finding actual ways of arbitrating the dispute, rather than writing the conflicting positions off as being simply incommensurable.

To begin tackling the issue, we need to lay down a certain amount of background information. For Kuhn, what qualifies as a paradigm, and how can we tell if we have conflicting or incommensurable paradigms? The distinctions are not likely to be "hard and fast", but certain common traits can be identified. The paradigm can be summarized as follows: it is a network of conceptual, theoretical, instrumental, and methodological commitments shared by a community of scientists. This doesn't mean to say that *all* the concepts, *all* the theories, etc. must be held by *all* the researchers working under a particular paradigm; it only means to say that there is great overlap of these entities within the paradigm. Additionally, there needs to be some agreement on the types of "puzzles" on which science should work, and some agreement on the types of answers which would satisfy those puzzles. There may be a certain amount of dissent within a paradigm, but it is not on the scale of radical disagreement over foundational principles or large-scale methodological assumptions. More likely, it will be at the level of the content of theories or experiments, rather than the structural form of the theories themselves.

In contrast to dissent from within a paradigm, Kuhn would view two communities to be functioning under different or rival paradigms if they differ in fundamental conceptual ideas concerning how to go about justifying theories,

what sorts of puzzles are appropriate problems, what sorts of answers would satisfy those puzzles, or other major principles that affect the very *practice* (rather than simply the empirical content) of science. Additionally, competing paradigms have the quality of incommensurability – that is, each paradigm uses its own rules and standards to justify its own position, thereby excluding the possibility of appealing to an agreed-upon objective standard. In Kuhnian science, there can be no objective standard – each scientist works within one paradigm and one paradigm only, meaning that the standards for comparison always are formulated with relation to one's fundamental assumptions inherited from that paradigm.

Whether or not we accept Kuhn's description of science as literally true, we can see some clear comparison between his description and the stand-off between the Bayesians and the error-statisticians in the philosophy of statistics. Clearly there are large-scale differences in both theory and methodology, and – while they have some puzzles in common – the type of answer which satisfies those puzzles and the type of justification required for those answers are obviously vastly different. But in order to get some practical mileage out of Kuhn, we cannot stop at simply identifying the paradigms in conflict. We need to also decipher what exactly the fundamental sources of the conflict are, and why they

appear to be incommensurable. Only once we have that laid out before us can we hope to find some way of navigating to a solution.

At the level of theory, there are clear differences between the Bayesian and error-statistical approaches. Perhaps the most fundamental of these is the question of the criteria for valid scientific inference. The error-statistician believes that justified scientific inference takes place when we have designed an experiment in which the probability of error is quite low with respect to our ability to detect the error. Methodologically, the error statistician wants to set the alpha probability of a Type I error (the error of rejecting a null hypothesis that is in fact true) to some very low value (such as .02), and to then calculate the beta probability of the Type II error (the error of accepting a false hypothesis) based upon the chosen alpha value. For a test with appropriate "power" (the ability to detect difference), the error-statistician can then also calculate the "severity" of the test – that is, the probability that we would have achieved the same experimental test result were the result in fact incorrect.

The Bayesian theory of statistical inference is radically different. The Bayesian either calculates or assigns some prior probability that the given hypothesis is true, and then recalculates (using Bayes' theorem) the

probability of that hypothesis in light of the received experimental evidence to produce a posterior probability of the truth of the hypothesis. Since the Bayesian includes prior probabilities (or prior "degrees of belief") in the experiment, the output is intended to serve as a "before and after" picture of our degree of belief in the hypothesis in light of the experimental evidence.

There are, of course, many intricate nuances and background assumptions to both statistical approaches. Beneath the surface of the Bayesian method lies an acceptance of the Likelihood Principle: namely, that if two experiments produce equivalent posterior probabilities in support of some hypothesis, that the evidence used in the two experiments is of the same experimental import (or are evidentially equivalent). Error statisticians reject this principle, arguing that it leads to the conclusion that stopping rules then become irrelevant, meaning that *any* hypothesis can become experimentally justified if we perform enough trials. The Bayesian replies that the error-statistical position is ridiculous, and that it leads to cases of contradictory test output using data of identical evidential import. The error-statistician replies that this is not at all ridiculous, but depends upon how well-tuned the tests were with regard to the error probabilities and test power. The debate continues ad

infinitum, with neither side willing to budge on the foundational assumptions that cascade up throughout the entirety of their statistical practice.

While the disagreement has all of the symptoms of a Kuhnian paradigm divide, the point is not to demonstrate that two points of view are or are not "truly" paradigmatic. To do so would be to imply that paradigms are some "thing" in the world, rather than just a philosopher's means of dividing up the world. Clearly there are statisticians who are not wholly committed to either side of the debate, those who pick and choose from the methods of each side as seems appropriate to solve particular problems, and those who simply don't believe that the difference is of practical significance. Saying that there are "actual" paradigms at work here matters little. What does matter is that it seems a useful way of talking about the problem. From *within* either camp, the rift certainly appears incommensurable. To accept the theories and assumptions of one is to reject those of the other. As Deborah Mayo puts it from the error-statistical perspective, "You cannot be just a little bit Bayesian" (Mayo 319.)

Does that fact that the split appears incommensurable from within either supposed paradigm mean that it is also incomparable to an observer who resides outside of either paradigm? Or would any standard by which to measure them

necessarily reside in some third paradigm, which Kuhn would consider just as subjective and value-laden as the first two? While I make a marked break from him here, I believe that Kuhn himself has unwittingly provided us with a possible answer. At the close of chapter four of *The Structure of Scientific Revolutions*, he says:

Finally, at a still higher level, there is a set of commitments without which no man is a scientist. The scientist must, for example, be concerned to understand the world and to extend the precision and scope with which it has been ordered. That commitment must, in turn, lead him to scrutinize, either for himself or through his colleagues, some aspect of nature in great empirical detail. And, if that scrutiny displays pockets of apparent disorder, then these must challenge him to a new refinement of his observational techniques or to a further refinement of his theories. Undoubtedly there are still other rules like these, ones which have held for scientists at all times (Kuhn 1996.)

The value of this passage should not be overlooked, for it provides us with a way out of the paradigm problem. Science (if it is to be considered science at all) must satisfy certain requirements, *regardless* of its background assumptions or foundational theories. It must scrutinize some aspect of nature in great detail, and it must be willing to refine its techniques and theories with regard to disorder in the experimental evidence. Failure to satisfy these criteria will disqualify *any* would-be science.

What bearing, then, does this have on the question of Bayesianism, error statistics, and scientific inference? Quite a bit, I think. If the debate centers solely upon

the quality of theories, then it will indeed be damned to eternal gridlock. But to do so would be to forget the purpose of doing science in the first place. The goal of science is *not* to produce elegant or even consistent theories. The goal of science is rather to provide adequate descriptions and reliable predictions about things in the world – about some aspect of nature. While in most cases a "tidy" theoretical framework will be able to satisfy this criterion more adequately than one infested with logical problems, there is no reason to believe that this is *necessarily* the case. Given a choice of two paradigms – one with elegant theories but a poor predictive track record, and one with slapdash theories but consistent and precise predictive accuracy – I would think that science would want to privilege the one that gives us greater accuracy about the world, regardless of the way it looks when represented symbolically.

And so I propose a re-framing of the debate over statistical inference. This re-framing moves the discussion away from the consistency of theories, which can bear no fruit if allowed to exist in a rhetorical vacuum. Despite profound theoretical differences, the Bayesian and error-statistical approaches do share certain puzzles in common, although they disagree on the way to arrive at answers. Some of those puzzles are bound to be empirical – puzzles with experimentally

verifiable results. It is the ability to provide descriptive and predictive accuracy in those puzzles which should serve as the benchmark for evaluating the theoretical and methodological claims of either party. Obviously (in order to remain scientific), theories will be adjusted in order to better describe the empirical data. But those theoretical adjustments cannot themselves be the basis for comparison. It is the ability of those revised theories to *do work* that will provide us with information about their scientific value.

At this point, certain objections will surely be raised. Foremost of these will be the objection that if empirical adequacy were the only issue at stake, the debate would never have arisen in the first place. The cases in which we utilize statistics are ones in which we don't have direct access to all of the empirical data. Therefore, appealing to accountability to the data for verification of the method gets us nowhere – if we knew what the data *said*, we wouldn't be doing a statistical analysis in the first place.

Obviously, it would be foolish to deny that this is entirely true for most cases of statistical inference. The very point of inference is to evaluate what we do know, and then to make extrapolations to what we don't know. Statistical sampling embodies this principle better than nearly anything else. However, I believe

that to say that the conclusions of statistical experiments are *never* verifiable in the long run would be to oversimplify. While many (or perhaps most) statistical conclusions attempt to provide descriptions of past occurrences or current trends, others are used to warrant very specific predictions about future events. And of that subset of predictive experiments, some of the predictions are easily verifiable – perhaps not in the short term, but perhaps in the long run. It is those predictions which I propose to use as a litmus test for the others, in order to determine the reliability and accuracy of a theoretical framework in describing things in the world.

The next obvious objection will be that to do so, we have merely moved the problem of justified inference up a level. At first glance, this objection does indeed seem to create a problem. In effect, we intend to take a measurable sample of all statistical experiments, to perform a second-order experiment on them, and then to use the conclusion to make an inference about statistical approaches as a whole. Will we view that second-order experiment through Bayesian or error-statistical lenses? Aren't we right back where we started?

Perhaps not. If we had to attempt evaluations of Bayesianism and error statistics independently to provide a degree of confirmation for each, then we certainly do

have a problem. But in this case, degree of confirmation is not actually what we're after. What we're after is an empirical comparison between two methods. To create a simplistic model: if we do a long-run series of empirically verifiable predictive experiments employing both the Bayesian and error-statistical approaches to the problem, and award one "point" per problem to the approach that provides the more accurate answer, then if there is indeed substantial theoretical advantage to one approach over the other, it will manifest itself in a point difference. I don't mean to suggest that the comparison would be quite so simple to perform in practice, but it does seem like the type of comparison that would provide useful information toward substantiating the seemingly incommensurable theoretical claims.

At this point the error statistician will protest that we still have a potential stopping rule problem — the Bayesians will want to continue running tests until they manage to gain an advantage. This, too, appears very sticky at first glance. If we assign a set number of iterations, then we concede the error statistician's point that the Likelihood Principle contains a problem. If we set no limit to the number of iterations, then we implicitly accept the Likelihood Principle, and thus the entire Bayesian framework founded upon it. The solution that I propose in fact appeals to neither: what if we set

no terminal point for the experiment in question, but also accept no emergent point as terminal, either? In effect, the experiment is endless and never conclusive. The Bayesian cannot choose to halt the experiment when they achieve advantage, and neither can they complain that we've conceded to the error-statistical critique of the Likelihood Principle. Instead of a conclusive experiment, we simply provide "status reports" at set intervals.

To qualify these admittedly simplistic ideas: I don't intend for my proposed "experiment" idea to be taken literally. I simply mean to demonstrate the fact that there is a shared goal for both sides of the argument at hand, and that goal is to provide adequate inference about the world around us. Certain types of statistical predictions can be verified in the long run, and I don't find logical inconsistency in using the accuracy and precision of those predictions as a basis of comparison for two otherwise incommensurable methodologies. I believe that we can borrow from Kuhn for as long as it proves fruitful for making distinctions, after which we must abandon incommensurability in favor of comparison and accountability to the empirical facts of the matter. At that point, incommensurability gives way to a weaker incompatibility, and that incompatibility may be one which can be resolved — not through comparison of theories, but through comparison of the abilities of

those theories to accurately and precisely predict observable events in the world.

References

- Kuhn, Thomas S. *The Structure of Scientific Revolutions* (3rd ed.) 1996. The University of Chicago Press: Chicago.
- Mayo, Deborah G. *Error and the Growth of Experimental Knowledge*. 1996. The University of Chicago Press: Chicago.