RMM Vol. 2, 2011, 103–114
Special Topic: Statistical Science and Philosophy of Science
Edited by Deborah G. Mayo, Aris Spanos and Kent W. Staley
http://www.rmm-journal.de/

Sir David Cox and Deborah Mayo

A Statistical Scientist Meets a Philosopher of Science: A Conversation between Sir David Cox and Deborah Mayo (as recorded, June, 2011)

COX: Deborah, in some fields foundations do not seem very important, but we both think foundations of statistical inference are important; why do you think that is?

MAYO: I think because they ask about fundamental questions of evidence, inference, and probability. I don't think that foundations of different fields are all alike; because in statistics we're so intimately connected to the scientific interest in learning about the world, we invariably cross into philosophical questions about empirical knowledge and inductive inference.

COX: One aspect of it is that it forces us to say what it is that we really want to know when we analyze a situation statistically. Do we want to put in a lot of information external to the data, or as little as possible. It forces us to think about questions of that sort.

MAYO: But key questions, I think, are not so much a matter of putting in a lot or a little information. Default Bayesians might say 'we don't make you put in more than the frequentist (just give us the model and data, we do the rest)'. What matters is the kind of information, and how to use it to learn. This gets to the question of how we manage to be so successful in learning about the world, despite knowledge gaps, uncertainties and errors. To me that's one of the deepest questions and it's the main one I care about. I don't think a (deductive) Bayesian computation can adequately answer it.

COX: It's also an issue of whether one looks to foundations just to provide a basis for what one does, and to enable us to do things a bit better, or whether it is to provide a justification of what we do; that's a bit different. Does one learn about the world because foundations of statistics are sound?

MAYO: No, but sound foundations are relevant to success in learning, at least if statistics is seen as formalizing lessons for how we deliberately avoid being led astray due to limited information and variability. That's my view; it may not

be shared by anyone, I realize. Seriously, I've had people say that foundations of statistics can't be relevant to learning in general because in the 17th or 18th centuries, say, scientists were learning even without modern statistics.

COX: Of course that's historically slightly misleading.

MAYO: My point is that when they were reasoning and learning they were doing something akin to reasoning in statistics.

COX: The kind of science that they were mostly doing did not call for elaborate statistical analysis but if it did, as in some astronomical problems, they would use statistics.

MAYO: Yes, it's as if there was still low hanging fruit not calling for explicit statistics.

COX: Something like that.

COX: There's a lot of talk about what used to be called inverse probability and is now called Bayesian theory. That represents at least two extremely different approaches. How do you see the two? Do you see them as part of a single whole? Or as very different?

MAYO: It's hard to give a single answer, because of a degree of schizophrenia among many Bayesians. On paper at least, the subjective Bayesian and the so-called default Bayesians, or whatever they want to call themselves,[1] are wildly different. For the former the prior represents your beliefs apart from the data, where disagreement is accepted and expected, where at most there is long-run convergence. Default Bayesians, by contrast, look up 'reference' priors that do not represent beliefs and might not even be probabilities, but give set rules to follow. Yet in reality default Bayesians seem to want it both ways. They say: 'All I'm trying to do is give you a prior to use if you don't know anything. But of course if you do have prior information, by all means, put it in.' It's an exercise that lets them claim to be objective, while inviting you to put in degrees of belief, if you have them. The prior, they like to say, gives a 'reference' to compare with your subjective prior, but a reference for what?

COX: Yes. Fisher's resolution of this issue in the context of the design of experiments was essentially that in designing an experiment you do have all sorts of prior information, and you use that to set up a good experimental design. Then

[1] 'Objective' Bayesian has caught on, but it is used in a way that seems at odds with objectivity in science. Granted, following a formal stipulation or convention is free from anything personal, but how does being impersonal in this sense promote the goal of using data to distinguish correct from incorrect claims about the world? Scientific objectivity, it seems to me, concerns the latter.

when you come to analyze it, you do not use the prior information. In fact you have very clever ways of making sure that your analysis is valid even if the prior information is totally wrong. If you use the wrong prior information you just got an inefficient design, that's all.

MAYO: What kind of prior, not prior probability?

COX: No, prior information, for example, a belief that certain situations are likely to give similar outcomes, or a belief that studying this effect is likely to be interesting. There would be informal reasons as to why that is the case that would come into the design, but it does not play any part in the analysis, in his view, and I think that is, on the whole, a very sound approach. Prior information is always there. It might be totally wrong but the investigator must believe something otherwise he or she he wouldn't be studying the issue in the first place.

MAYO: Insofar as the background influences the choice of model, the ultimate inference does seem to be influenced by it.

COX: It didn't influence the choice of model so much as it influenced the design of the experiment. The analysis of the experiment is independent of the prior information.

MAYO: Yes, but clearly the model could be wrong. Did Fisher talk about testing models?

COX: Yes, as with much of his work, it's a bit difficult to get at, but he gave an argument based on his development of the idea of sufficiency and most of the emphasis is on the idea that the sufficient statistics tell you about the parameters in the model. But he also pointed out that the conditional distribution given the sufficient statistic would provide good ways of testing the model. He wrote a very nice paper on this, quite late in his life actually. But the key issue, isn't it, is whether, when you think about objective and personalistic Bayesians, are they really trying to do such very different things that they are to be treated as totally different approaches, even if from a formal mathematical point of view they might look the same.

MAYO: Well, as I say, the default Bayesian vacillates or wants it both ways (depending on the audience): In theory the two are doing something radically different, but in practice the default Bayesians often seem to be giving manuals for reference priors as a stop-gap measure to be replaced by degrees of belief, when you get them. Default Bayesians, some of them, admit to being opportunistic, wanting to keep their foot in the door lest they be ignored by scientists who oppose doing a subjective analysis. But at the same time they can be found denying the reference prior is to be used for inference, but only to 'calibrate' (in

a way they don't justify) your subjective priors. It's hard to see how one could really criticize the result; it could be due to subjective opinions or your favorite manual of default priors.

COX: Yes, but what they do is a different issue from their conceptual theory and it seems to me that their conceptual theories are trying to do two entirely different things. One is trying to extract information from the data, while the other, personalistic theory, is trying to indicate what you should believe, with regard to information from the data and other, prior, information treated equally seriously. These are two very different things.

MAYO: Yes, except that I have a real worry even with respect to the claim that default Bayesians are about learning from the data. The whole foundational issue is what does that mean? How do you learn from data? For them it might be that the data enter, I don't know, perhaps through a report of likelihoods. So I question even how they implement the goal of learning from the data. To me, I cannot scrutinize what I've learned from the data without an error probabilistic analysis, and insofar as default Bayesians are saying they don't do that, even to distinguish between poor and good inferences, then to me they're not finding out what's in the data. They seem to think the likelihood function tells you what's in the data and I'm saying that's not enough.

COX: Well the primitive idea is that there is a model, and data, and the data are either consistent with the model in some reasonable sense, or inconsistent with the model, and the job of statistics is to put that dichotomy on a more settled ground.

MAYO: OK, but statistical methods don't all assess this in the same way. The very idea of a consistent fit is ambiguous.

COX: Yes. No model is going to describe all aspects of a realistic set of data, it's too complicated. It has got to describe those features which in some sense matter. That's a difficult issue of the link between theory and application, more than an issue of principle.

MAYO: Yes, but statistics should give reliable ways to assess consistency, at least between statistical hypotheses and data.

COX: There are situations where it is very clear that whatever a scientist or statistician might do privately in looking at data, when they present their information to the public or government department or whatever, they should absolutely not use prior information, because the prior opinions on some of these

prickly issues of public policy can often be highly contentious with different people with strong and very conflicting views.

MAYO: But they should use existing knowledge.

COX: Knowledge yes. Prior knowledge will go into constructing the model in the first place or even asking the question or even finding it at all interesting. It's not evidence that should be used if let's say a group of surgeons claim we are very, very strongly convinced, maybe to probability 0.99, that this surgical procedure works and is good for patients, without inquiring where the 0.99 came from. It's a very dangerous line of argument. But not unknown.

MAYO: (laughs).

COX: Similar issues arise in public policy on education or criminology, or things like that. There are often very strong opinions expressed that if converted into prior probabilities would give different people very high prior probabilities to conflicting claims. That's precisely what the scientist doesn't want.

MAYO: Yes, I agree. I don't know how they get away with saying in reputable Bayesian texts, often, things like: there's an objective frequentist account and then there's a Bayesian account that deals with decisions and utilities. The latter is more relevant, they allege, since it tells you what decisions to make, and you obviously want to make decisions. They don't question whether they can first get a reliable evidential basis for decisions, yet it's used as a selling point.

COX: Yes. Well the decision theory aspect is important, isn't it, because many investigations are intended at some point to influence a decision, about how patients are treated or if a policy on education should be followed, or whatever, but that's very different from presenting the information or determining what it is reasonable to believe in the light of the data. How does a philosopher see that?

MAYO: Well I take issue with most philosophers insofar as they assume 'rationality' is a matter of (Bayesian) decision-making based on prior beliefs. I am also at odds with those who seem to hold that what makes an account of evidence relevant for 'epistemology' is that it's framed in terms of an agent's beliefs. But that is generally just an analytical exercise, whereas philosophers should really favor accounts that tell us how to arrive at reliable inferences and well-tested claims (or 'beliefs' if one insists)! Since Kuhn, many infer from failed analytic attempts that we can only give menus of properties that at different times and contexts scientists would like evidence to have (consistency, scope, simplicity, reliability, etc.). As for decision and inference, I think they should be distinct, and I'm really glad that you highlight that, because I am often beaten up on this. Many question how there can even be a difference (between what is the case and what you should do). By and large, those who doubt the very idea that there

can be a difference embrace some kind of relativism or social constructivism in philosophy. But even some philosophers who claim not to be subjectivists insist, to my bewilderment, that an account of inference should itself be an account of decision making, combining learning goals with other kinds of losses and values. One is free to call any inference a kind of decision, of course, but the 'utilities' would have to reflect the goal of finding things out correctly; but then embedding it in a decision framework doesn't help, but it hides a lot.

COX: I have often been connected with government decision-making. The idea that we would present people's opinions unbacked by evidence would have been treated as ludicrous. We were there as scientists to supposedly provide objective information about the issue. Of course I know there is difficulty with the idea of total objectivity but at least it should connect with truth, to the goal of getting it right.

MAYO: The evidential report should be constrained by the world, by what is actually the case.

COX: Yes.

MAYO: I do find it striking that people could say with a straight face that we frequentists are not allowed to use any background information in using our methods. I have asked them to show me a book that says that, but they have not produced any. I don't know if this is another one of those secrets shared only by the Bayesian Brotherhood.

COX: Well it's totally ridiculous isn't it.

MAYO: Then again, I suppose we don't see statistical texts remedying this in a way that makes it conspicuous, that acknowledges this criticism and emphasizes that frequentists never advocated doing inference from a blank slate, but that you need to put together pieces, combine other tests and well-probed hypotheses. (We emphasize this in Cox and Mayo 2010.)

COX: Yes, you have to look at all the evidence but the main purpose of statistical analysis is to clarify what it is reasonable to learn from the specific set of limited data. It is a limited objective. Would you agree?

MAYO: Yes, but I want to say more of what this means and, again, I would insist that I cannot know what it is reasonable to infer without knowing something about the method's ability to have demonstrated the mistake of relevance. I don't want to just know that this model beautifully fits the data, and would have predicted the data; the test may be totally lacking in severity. The severe tester insists on asking: but could you have unearthed some error, were it present (the stern taskmaster depicted in my conference slides)? What theories

could your data not have been able to rule out very well, if at all? This kind of self-scrutiny and self-correcting is an important source of progress, and is not given a home in standard epistemologies. It takes literally the idea of learning from error, and of developing new hypotheses by noting rival theories that would all be consistent with given data, at a given level.

COX: Now the notion that statistical methods should have good long-run properties is apparently ill at ease with the goal of reaching a good conclusion from this unique set of data under analysis, as Fisher sometimes put it. How do you see that issue?

MAYO: My idea is that at least in the case where we are interested in learning from this data set (in other cases we may not be), the hypothetical long-run properties of the method serve to tell us what mistakes this method would have, with fairly good probability, detected and which it wouldn't have. The sampling distribution depicts what would be expected were we wrong about some claim. Therefore I advocate using the long-run properties to scrutinize the ability or incapacity of the tools, and thereby reach an inference about some aspect of the procedure that generated this particular data. There is no conflict when the long-run properties are relevant to this purpose. I admit that this position isn't logically deducible from formal statistics, that is why it is a *philosophical* position, a philosophy of statistics, one that I claim makes sense of, and justifies, how we successfully find things out in science and in ordinary life, despite using error-prone tools. But I don't know if I'm being as clear as I would like to be.

COX: This is very counter to a literal interpretation of Neyman, isn't it? It's very different in fact.

MAYO: Yes, it's really puzzling because there are places where Neyman spoke this way (i.e., inferentially), and certainly Pearson did, often, even though he didn't speak enough, philosophically. Yet there are places where Neyman is just as behavioristic as you can be, and so he does deserve that label despite the places where Neyman speaks 'inferentially' and despite the fact that I know he was mostly drawing a contrast with the Bayesian view of inductive inference, and so he introduces a different term. Instead of telling us how to adjust our beliefs, statistical methods become tools to adjust our behavior in the face of limited information. That's an interesting idea of his, but he went overboard. Sometimes he even extolled the fact that lumping together, not just different studies of a given area but all scientific applications, lets us show that, on the whole, we are wrong with low probability. But I think he was just carried away with these results that were really neat (following from the law of large numbers).

COX: It is relevant that Egon Pearson had a very strong interest in industrial design and quality control.

MAYO: Yes, that's surprising, given his evidential leanings and his apparent distaste for Neyman's behavioristic stance. I only discovered that around 10 years ago; he wrote a small book.[2]

COX: He also wrote a very big book, but all copies were burned in one of the first air raids on London.

MAYO: What was the story again about this burning? You told me and Aris once.

COX: All of the copies in the warehouse where the books were stored were burned in one of the first air raids on London, in the summer of 1940.

COX: Anyway, I think the issue of making frequentist statements convincing for a particular set of data that one happens to have is a critical issue. I don't think it's too big a deal in practice but conceptually it's important.

MAYO: Yes, I want to suggest that that's how you should try to apply the frequency statements. Think about just an ordinary instrument: I want to know what its capabilities are, and that's the role of the long-run error probabilities. (You recall my favorite example of determining my weight gain by means of information of the precision and reliability of a group of scales.) I've always thought this was a useful twist on what Birnbaum thought (e.g., his *confidence concept*), and what all the other attempts at evidential interpretations (of Neyman-Pearson statistics) say. Maybe it is just a little twist, but I've increasingly found that it helps to solve key problems. The reasoning directs you to consider specific mistakes of relevance; and you want to evaluate the (formal) error probabilities of a method in relation to the (informal) errors of inference that the method ought to have been capable of informing us about, at least if there is to be a warrant for ruling out those mistakes. Unfortunately, many who try to 'reconcile' Bayesian and frequentist methods tend to appeal to the most radical behavioristic goals, claiming they don't do too badly in the asymptotic long-runs. This does not suffice for warranted inferences in the particular case. It's like they wind up with the worst of both worlds.

COX: It is sometimes claimed that there are logical inconsistencies in frequentist theory, in particular surrounding the strong Likelihood Principle (SLP). I know you have written about this, what is your view at the moment.

---

[2] I thank Aris Spanos for locating this work of Pearson's from 1935; and for his continued astuteness on matters both historical and mathematical.

MAYO: What contradiction?

COX: Well, that frequentist theory does not obey the strong Likelihood Principle.

MAYO: The fact that the frequentist rejects the strong LP is no contradiction.

COX: Of course, but the alleged contradiction is that from frequentist principles (sufficiency, conditionality) you should accept the strong LP. The (argument for) the strong LP has always seemed to me totally unconvincing, but the argument is still considered one of the most powerful arguments against the frequentist theory.

MAYO: Do you think so?

COX: Yes, it's a radical idea, if it were true.

MAYO: You're not asking me to discuss where Birnbaum goes wrong (are you)? [Of course Birnbaum himself rejected the strong LP because it prevented the control of error probabilities.]

COX: Where *did* Birnbaum go wrong? (Note that in his last paper he recommended confidence intervals.)

MAYO: I am not sure it can be talked through readily, even though in one sense it is simple; so I relegate it to an appendix. It turns out that the premises are inconsistent, so it is not surprising the result is an inconsistency. The argument is unsound: it is impossible for the premises to all be true at the same time. Alternatively, if one allows the premises to be true, the argument is not deductively valid. You can take your pick.

Appendix

a. Basics:

Even a sketch of the argument requires being clear on several notions; here I just define the basics of the Strong Likelihood Principle (SLP).

> *The strong LP is a conditional claim:*
> (SLP): If there are two experiments E' and E" with different probability models but with the same unknown parameter μ, and x' and x" are observed results from E' and E" respectively, where the likelihood of x' and x" are proportional to each other (i.e., differ only by a constant), then x' and x" ought to have the identical evidential import for any inference concerning parameter μ.

For instance, E' and E" might be Binomial sampling with n fixed, and Negative Binomial sampling, respectively. For a more extreme example, E' might be sampling from a Normal distribution with a fixed sample size n, and E" might be the corresponding experiment that uses this stop rule: keep sampling until you obtain a result 2-standard-deviations away from a null hypothesis.

Suppose we are testing the null hypothesis that $\mu = 0$. The SLP tells us that once you have observed a 2-standard-deviation result, there ought to be no evidential difference between its having arisen from experiment E', where n was fixed at 100, and experiment E" where one is allowed to stop at n = 100 (i.e., it just happens that a 2-standard-deviation result was observed after n = 100 trials).

The key point is that there is a difference in the corresponding p-values from E' and E", which we may write as p' and p", respectively. While p' would be ~ .05, p" would be much larger, ~ .3. The error probability accumulates because of the optional stopping. Clearly p' is not equal to p", so the two outcomes are not evidentially equivalent for a frequentist. This constitutes a violation of the strong LP (which of course is just what is proper for a frequentist). For a fuller discussion of this part, see Mayo 1996, chapters 9 and 10; Mayo and Kruse 2001.[3]

b. Birnbaum's Argument in a Nutshell:

The first step is to take any violation of the SLP, that is, a case where the antecedent of the LP holds, and the consequent does not hold. Assume then that the pair of outcomes x' and x", from E' and E" respectively, represent a violation of the SLP. We may call them *SLP pairs*.

Step 1:
Birnbaum will describe a funny kind of 'mixture' experiment based on an SLP pair; I call it a Birnbaum (BB) experiment. Within this mixture, it appears that we must treat x' as evidentially equivalent to its SLP pair, x".

In particular, having observed x' from the fixed sample size experiment E', I am to imagine x' resulted from getting heads on the toss of a fair coin, where tails would have meant performing E". Further, in the BB experiment, we are required to erase the fact that x' came from E'—the test statistic says, in effect, it could have come from either E' or E". Call this test statistic of a BB experiment: T-BB.

In reporting a p-value associated with x', for example, instead of reporting its p-value p', we are to report the average of p' and p": (p' + p")/2. The test statistic T-BB is sufficient, technically, but the argument overlooks that an error statistician still must take into account the sampling distribution. In this case it refers to the distribution of T-BB. That's what dooms the 'proof', as we see in Step 2:

[3] Many sources of the Birnbaum argument can be found, unsurprisingly with 'gaps' left as an exercise. See for example Casella and Berger 2002.

Step 2:
A second premise is now added to the argument. Here it is reasoned that once we know that x' came from experiment E', we should not treat it as the BB experiment (of Step 1), but rather we should report x' came from experiment E', and evaluate the result in the usual way. Now for a frequentist, the usual way would be to report p'. Likewise if we knew we had obtained x" from experiment E", we are to report p", according to Step 2.

It is actually difficult to even reformulate the argument as deductively valid, since if the premises are made true, the terms are forced to change within the argument. But it is interesting to try and do so, in order to highlight the confusion.

> *An abbreviation*:
> Let '*equiv*' be an abbreviation for 'is or should be evidentially equivalent to'.

To streamline the argument further, let us take the evidential assessment to be in terms of p-values. We have the following argument with (0), (1) and (2) as premises, (3) as conclusion:

(0) Let x' from E' and x" from E" be an arbitrary example of a pair that violates the SLP.
(1) In drawing inferences from outcomes in a Birnbaum experiment:
x" *equiv* x" (both of which equal (p' + p")/2)
(2) In drawing inferences from outcomes in a Birnbaum experiment
  (a) x' *equiv* p'
  (b) x" *equiv* p"

---

(3) Conclusion: from (1) and (2a and 2b): p' *equiv* p".

More generally, the conclusion would be for *any* SLP pair, x' and x", x' *equiv* x". We have thus put Birnbaum's argument into valid form.

But from (0), we know x' and x" form a SLP violation, so, from (0) not (p' *equiv* p"). Thus it would appear the frequentist is led into a contradiction.

The problem is that in order to infer the conclusion, p' *equiv* p", the premises of the argument must be true, and it is impossible to have premises (1) and (2) true at the same time. Premise (1) is true only if we use the sampling distribution given in the Birnbaum experiment (averaging over the SLP pairs). This is the sampling distribution of T-BB.

Yet to draw inferences using this sampling distribution renders both (2a) and (2b) false. Their truth requires 'conditioning' on the experiment actually performed, or rather, they require we not 'Birnbaumize' the experiment from which the observed LP pair is known to have actually come!

Although I have allowed premise (1) for the sake of argument, the very idea is extremely far-fetched and unmotivated.[4] It is worth noting that Birnbaum himself rejected the SLP (Birnbaum 1969, 128): "Thus it seems that the likelihood concept cannot be construed so as to allow useful appraisal, and thereby possible control, of erroneous interpretations." For further discussion on this part, see Mayo 2010; Cox and Mayo 2010.

References

Birnbaum, A. (1969), "Concepts of Statistical Evidence", in: Morgenbesser S., P. Suppes and M. White (eds.), *Philosophy, Science, and Method: Essays in Honor of Earnest Nagel*, New York: St. Martin's Press, 112–143.

Casella, G. and R. Berger (2002), *Statistical Inference*, 2nd ed., Pacific Grove: Duxbury.

Cox, D. (1978), "Foundations of Statistical Inference: The Case For Eclecticism", *Australian Journal of Statistics* 20(1), 43–59.

— and D. Mayo (2010), "Objectivity and Conditionality in Frequentist Inference", in: Mayo and Spanos 2010, 276–304.

Mayo, D. (1996), *Error and the Growth of Experimental Knowledge*, Chicago: University of Chicago Press.

— (2010), "An Error in the Argument from Conditionality and Sufficiency to the Likelihood Principle", in: Mayo and Spanos 2010, 305–314.

— and D. Cox (2006), "Frequentist Statistics as a Theory of Inductive Inference", in: Rojo, J. (ed.), *Optimality: The Second Erich L. Lehmann Symposium*, vol. 49 of Lecture Notes-Monograph Series, Institute of Mathematical Statistics (IMS), Beachwood, 77–97. Reprinted in Mayo and Spanos 2010, 247–275.

— and M. Kruse (2001), "Principles of Inference and Their Consequences", in: Corfield, D. and J. Williamson (eds.), *Foundations of Bayesianism*, Dordrecht: Kluwer, 381–403.

— and A. Spanos (2010) (eds.), *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, Cambridge: Cambridge University Press.

— and D. Cox (2010), "Frequentist Statistics as a Theory of Inductive Inference", in: Mayo and Spanos 2010, as reprinted from Mayo and Cox 2006, 247–275.

Pearson, E. S. (1935), *The Application of Statistical Methods to Industrial Standardization and Quality Control*, London: British Standards Institution.

[4] For example, one has to first observe the outcome, then, if it happens to be an outcome with an SLP pair, construct a mixture that could have produced it, and then analyze the result as if it had actually resulted from a mixed experiment (even though it did not), and proceed as if one had planned all along to average over the mixture (erasing which experiment it came from). As Cox (1978, 54) notes, the argument would require considering all pairs that could arise as what I call SLP pairs. Pre-data, it would seem to require averaging over all of the hypothetical possibilities. If the outcome does not have an SLP pair, the Birnbaum argument instructs you to report it in the usual way, with no averaging. Ironically, then, only outcomes that do not have SLP pairs are treated in a way such that conditioning is correctly applied.