III An Error in the Argument from Conditionality and Sufficiency to the Likelihood Principle

Deborah G. Mayo

Cox and Mayo (7(II), this volume) make the following bold assertion:

> It is not uncommon to see statistics texts argue that in frequentist theory one is faced with the following dilemma, either to deny the appropriateness of conditioning on the precision of the tool chosen by the toss of a coin, or else to embrace the strong likelihood principle which entails that frequentist sampling distributions are irrelevant to inference once the data are obtained. This is a false dilemma: Conditioning is warranted in achieving objective frequentist goals, and the conditionality principle coupled with sufficiency does not entail the strong likelihood principle. The "dilemma" argument is therefore an illusion.

Given how widespread is the presumption that the (weak) conditionality principle (CP) plus the sufficiency principle (SP) entails (and is entailed by) the (strong) likelihood principle (LP), and given the dire consequence for error statistics that follows from assuming it is true, some justification for our dismissal is warranted. The discussion of the three principles (in 7(II)) sets the stage for doing this. The argument purporting to show that CP + SP entails LP was first given by Birnbaum (1962), although the most familiar version is that found in Berger and Wolpert (1988). To have a statement of the SLP in front of us for this discussion I restate it here:

> *The Strong Likelihood Principle (SLP)*: Suppose that we have *two* experiments E' and E'', with different probability models $f'_{Y'}(\boldsymbol{y}'; \theta)$ and $f''_{Y''}(\boldsymbol{y}''; \theta)$, respectively, with the same unknown parameter θ. If $\boldsymbol{y}'^*$ and $\boldsymbol{y}''^*$ are observed data from E' and E'', respectively, where the likelihoods of $\boldsymbol{y}'^*$ and $\boldsymbol{y}''^*$ are proportional, *then* $\boldsymbol{y}'^*$ and $\boldsymbol{y}''^*$ have the identical evidential import for any inference about θ.

This notation, commonly used in discussing this result, suggests a useful shorthand: we may dub those outcomes y'^* and y''^* that satisfy the requirements of the SLP "star pairs" from the experiments E' and E''.

Principles of inference based on the sampling distribution do not in general satisfy the (strong) likelihood principle; indeed, were the SLP to be accepted it would render sampling distributions irrelevant for inference once the data were in hand. Understandably, therefore, statisticians find it surprising that Birnbaum purports to show the SLP is entailed by the two principles accepted in most formulations of frequentist theory. (The SLP is, at least formally, incorporated in most formulations of Bayesian and purely likelihood-based accounts of statistical inference.)

Example 5 (p. 286) showed how frequentist theory violates the SLP; the following case is even more dramatic.

Example: Fixed versus sequential sampling. Suppose $\boldsymbol{Y}'$ and $\boldsymbol{Y}''$ are sets of independent observations from $N(\theta, \sigma^2)$, with σ known, and p-values are to be calculated for the null hypothesis $\theta = 0$. In test E' the sample size is fixed, whereas in E'' the sampling rule is to continue sampling until $1.96\sigma_y$ is attained or exceeded, where $\sigma_y = \sigma/n^{.5}$. Suppose E'' is first able to stop with n_0 trials. Then $\boldsymbol{y}''$ has a proportional likelihood to a result that could have occurred from E', where n_0 was fixed in advance, and it happened that after n_0 trials $\boldsymbol{y}'$ was $1.96\sigma_y$ from zero. Although the corresponding p-values would be different, the two results would be inferentially equivalent according to the SLP. According to the SLP, the fact that our subject planned to persist until he got the desired success rate – the fact that he *tried and tried again* – can make *no* difference to the evidential import of the data: the data should be interpreted in just the same way as if the number of trials was fixed at the start and statistical significance resulted. (Bayesians call this the stopping rule principle.) For those of us wishing to avoid misleading inferences with high or maximal probability, this example, "taken in the context of examining consistency with $\theta = 0$, is enough to refute the strong likelihood principle" (Cox, 1977, p. 54) because, with probability 1, it will stop with a "nominally" significant result even though $\theta = 0$.

By contrast, Savage declares:

> The persistent experimenter can arrive at data that nominally reject any null hypothesis at any significance level, when the null hypothesis is in fact true. . . . These truths are usually misinterpreted to suggest that the data of such a persistent experimenter are worthless or at least need special interpretation . . . The likelihood principle, however, affirms that the experimenter's intention to persist does not change the import of his experience. (Savage, 1962a, p. 18)

Before anyone came forward with independent grounds for accepting the SLP, one merely had a radically different perspective on the goals for sensible

accounts of statistics. This situation was to change dramatically with Birnbaum's argument that purported to provide just such grounds, using only premises apparently supported by frequentists. Doubtless this is why Savage (1962b) greeted Birnbaum's argument for the SLP as a "breakthrough." As it is often said, "the proof is surprisingly simple" (p. 467). However, it is precisely the overly quick series of equivalences so characteristic of each presentation that disguises how the vital terms shift throughout. Although important caveats to the Birnbaum argument have often been raised, the soundness of his results has generally been accepted by frequentists, likelihoodists, and Bayesians alike (see Casella and Berger, 2002; Lee, 2004; Lehmann, 1981; Robins and Wasserman, 2000; Royall, 1997). Yet a close look at the argument reveals it to be deeply (and interestingly) flawed.

At the risk of belaboring key elements of the Birnbaum experiment, I first informally outline the ingredients. Then the more usual notation can be grasped without getting dizzy (hopefully).

1 Introduction to the Birnbaum Experiment: E-BB

Let E' and E'' be expements some of whose outcomes are "star pairs," i.e., some y'^* from E' and y''^* from E'' satisfy the proportional likelihood requirement in the SLP. An example would be E', the normal test with n fixed, the significance level set to .05 and E'' the corresponding test with the rule: stop when a 0.05 statistically significant result is achieved. There are two steps to the Birnbaum experiment.

Step 1: Use E' and E'' to define a special type of mixture experiment: Flip a coin to determine whether E' or E'' is to be performed. One can indicate which is performed using statistic J: in particular, $j = 1$ and $j = 2$ when E' and E'' are performed, respectively. *So far this is an ordinary mixture.*

Step 2: Once you toss the coin and perform the indicated experiment, you are to report the result in the following way:

If the result has a star pair in the experiment not performed, then report it came from E', whether or not it came from E' or E''. Say $j = 2$ so the optional stopping experiment is run, and let the experiment achieve .05 significance when $n = 100$. Does this outcome have a "star pair" in experiment E'? Yes, its star pair is the case where n was fixed at 100, and statistical significance at the .05 level was achieved. The Birnbaum experiment instructs you to report: (E', y'^*) that the result came from the fixed sample size test, even though the result actually came from E''. We can construe his rule about

reporting results in terms of a statistic, call it T_{BB}, that erases the fact that a result came from E'' and writes down E' (in the case of starred outcomes):

$$T_{BB}(E'', y''^*) = (E', y'^*).$$

Let us abbreviate the special mixture experiment of Birnbaum's as E-BB. When (E', y'^*) is reported from E-BB, you know that the result could have come from E' or E'', but you do not know which.

We have not said what E-BB stipulates if the result does not have a star pair. In that case, report the results as usual. For example, let $j = 1$, so E' is performed, and say the results after the fixed number of trials is not statistically significant. Since these latter (case 2) results do not enter into the main proof we can focus just on the star pairs (case 1).

I have said nothing so far about the argument purporting to show the SLP, I am just describing the set-up. In setting out the argument, it will be useful to employ the abbreviation introduced in 7(II), p. 287:

Infr$_E(y)$ is the inference from outcome y from experiment E *according to the methodology of inference being used.*

Although this abbreviation readily lends itself for use across schools of inference, here we use it to discuss an argument that purports to be relevant for frequentist sampling theory; therefore, the relativity to the sampling distribution of any associated statistic is fundamental. Birnbaum is free to stipulate his rule for reporting results, so long as inferences are computed using the sampling distributions corresponding to T_{BB} in experiment E-BB. If this is done consistently however, the desired result (SLP) no longer follows – or so I will argue.

2 The Argument from the Birnbaum Experiment

I first trace out the standard version of the argument that purports to derive the SLP from sufficiency and conditionality (found, for example, in Berger and Wolpert, 1988; Casella and Berger, 2002). It begins with the antecedent of the SLP.

A. We are to consider a pair of experiments E' and E'' with f' differing from f'' where E' and E'' have some outcomes y'^* and y''^* with proportional likelihoods.

We may think of E' as the normal test with n fixed, and E'' the normal test with optional stopping as mentioned earlier. The argument purports to

show the following:

$$\text{Infr}_{E'}(y'^{*}) = \text{Infr}_{E''}(y''^{*}).$$

B. We are to consider a mixture experiment whereby a fair coin is flipped and "heads" leads to performing E' and reporting the outcome y', whereas "tails" leads to E'' and reporting the outcome y''. Each outcome would have two components (E^j, y^j) $(j = 0, 1)$, and the distribution for the mixture would be sampled over the distinct sample spaces of E' and E''.

In the special version of this experiment that Birnbaum puts forward, E_{BB}, whenever an outcome from E'' has a "star pair" in E' we report it as (E', Y'^{*}), (case 1), else we are to report the experiment performed and the outcome (case 2).

That is, Birnbaum's experiment, E_{BB}, is based on the statistic T_{BB}.

$$T_{\text{BB}}(E^j, y^j) = \begin{cases} (E', y'^{*}) \text{ if } j = 1 \text{ and } y' = y'^{*} \text{ or if } j = 2 \text{ and } y'' = y''^{*} \\ (E^j, y^j) \end{cases}$$

For example, if the result is (E'', y''^{*}), we are to report (E', y'^{*}).

Because the argument for the SLP is dependent on case 1 outcomes, we may focus only on them for now. (Note, however, that only case 2 outcomes describe the ordinary mixture experiment.) Now for the premises:

(1) The next step is to argue that in drawing inferences from outcomes in experiment E_{BB} the inference from (E'', y''^{*}) is or should be the same as if the result were (E', y'^{*}), as they both yield the identical output (E', y'^{*}) according to (sufficient) statistic T_{BB}. This gives the first premise of the argument:

$$(1)\ \text{Infr}_{E\text{-BB}}(E', y'^{*}) = \text{Infr}_{E\text{-BB}}(E'', y''^{*}).$$

(2) The argument next points out that WCP tells us that, once it is known which of E' or E'' produced the outcome, we should compute the inference just as if it was known all along that E^j was going to be performed. Applying WCP to Birnbaum's mixture gives premise (2):

$$(2)\ \text{Infr}_{E\text{-BB}}(\text{E}^{\text{j}}, \text{y}^{\text{j}*}) = \text{Infr}_{\text{Ej}}(\text{y}^{\text{j}*}).$$

Premises (1) and (2) entail the inference from y'^{*} is (or ought to be) identical to the inference from y''^{*}, which is the SLP.

Here I have written the argument to be formally valid; the trouble is that the premises cannot both be true. If one interprets premise (1) so that it comes out true, then premise (2) comes out false, and vice versa. Premise

(2) asserts: The inference from the outcome (E^j, y^{j*}) computed using the sampling distribution of E_{BB} is appropriately identified with an inference from outcome y^{j*} based on the sampling distribution of E^j, which is clearly false. The sampling distribution to arrive at $\text{Infr}_{E\text{-BB}}$ would be the convex combination averaged over the two ways that y^{j*} could have occurred. This differs from the sampling distributions of both $\text{Infr}_{E'}(\boldsymbol{y}'^*)$ and $\text{Infr}_{E''}(\boldsymbol{y}''^*)$.

3 Second Variation of the Birnbaum Argument

To help bring out the flaw in the argument, let us consider it in a different, equivalent, manner. We construe the premises so that they all are true, after which we can see the argument is formally invalid.

We retain the first premise as in the standard argument. To flesh it out in words, it asserts the following:

(1) If the inference from the outcome of the mixture, (E^j, y^{j*}), is computed using the unconditional formulation of E_{BB}, then the inference from (E', y'^*) is the same as that from (E'', y''^*).

$$(1)\ \text{Infr}_{E\text{-BB}}(E', y''^*) = \text{Infr}_{E\text{-BB}}(E'', y''^*).$$

We need to replace the false premise (2) of the standard formulation with a true premise (2)′:

(2)′ If the inference is to be conditional on the experiment actually performed then the inference from (E^j, y^{j*}) should be the same as $\text{Infr}_{E^j}(y^{j*})$.

Note we can speak of an inference from an outcome of E_{BB} without saying the inference is based on the sampling distribution of E_{BB} – avoiding the assertion that renders premise (2) of the standard variation false.

We have from the WCP: Once (E^j, y^{j*}) is known to have occurred, the inference should condition on the experiment actually performed, E^j.
The conclusion is the SLP:

Therefore, the inference from y'^* is or should be the same as that from y''^*.

Because our goal here is to give the most generous interpretation that makes the premises true, we may assume it is intended for the phrase "once (E^j, y^j) is known to have occurred" is understood to apply to premise (1) as well. The problem now is that, even by allowing all the premises to be

true, the conclusion does not follow. The conclusion could "follow" only if it is assumed that you both should and should not use the conditional formulation. The antecedent of premise (1) is the denial of the antecedent of premise (2)′. The argument is invalid; if made valid, as in the first rendition, it requires adding a contradiction as a premise. Then it is unsound.

4 An Explicit Counterexample

Because the argument is bound to seem convoluted for those not entrenched in the treatments of this example, it is illuminating to engage in the logician's technique of proving invalidity. We consider a specific example of an SLP violation and show that no contradiction results in adhering to both WCP and SP.

Suppose optional stopping experiment E'' is performed and obtains the 1.96 standard deviation difference when $n = 100$. Let σ be 1, $1.96\sigma_y = .196$. This value is sometimes said to be a "nominally" .05 significant result.

The calculation of the "actual" p-value would need to account for the optional stopping, resulting in a higher p-value than if the .196 had been obtained from a fixed-sample-size experiment. This leads to a violation of the SLP.

So we have

$$\text{Infr}_{E'}(\text{y}'^{*} = .196) = .05,$$

$$\text{Infr}_{E''}(y''^{*} = .196) = .37 \text{ (approximately)}.$$

Although these two outcomes have proportional likelihoods, they lead to different inferences. This result gives a clear violation of the SLP because $.05 \neq .37$.

$$\text{Infr}_{E'}(y'^{*} = .196) \neq \text{Infr}_{E''}(\text{y}''^{*} = .196).$$

Now the Birnbaum argument alleges that, if we allow this violation of the SLP, then we cannot also adhere to the WCP and the SP, on pain of contradiction (Berger and Wolpert 1988, p. 28). But we can see that no contradiction results from not-SLP and WCP and SP.

Sufficiency Principle (SP)

Where is sufficiency to enter? We know that SP applies to a single experiment, whereas E' and E'' are distinct experiments. But a mixture of the two

permits inference in terms of outcomes whose first component indicates the experiment run and the second the outcome from that experiment: (E^j, y^j). The mixture can have a single sampling distribution composed of the convex combination of the two.

For the mixed experiment we are discussing, the statistic associated with the Birnbaum experiment E_{BB} is T_{BB}, where

$$T_{\text{BB}}(E'', 1.96\sigma_y) = T_{\text{BB}}(E', 1.96\sigma_y) = (E', 1.96\sigma_y).$$

The unconditional *p*-value associated with the outcome from the mixed experiment would be based on the convex combination of the two experiments; it would be .5(.05 + .37) = .21. If we are to base inferences on the sampling distribution of T_{BB}, then in both cases the average *p*-value associated with .196 would be .21.

$$(1)\ \ \text{Infr}_{E\text{-BB}}(E'', .196) = \text{Infr}_{E\text{-BB}}(E',\ .196) = .21.$$

This may be seen as an application of the SP within E_{BB}. Sufficiency reports on a mathematical equivalence that results, provided that it is given that inference is to be based on a certain statistic T_{BB} (and its sampling distribution) defined on experiment E_{BB}. Having defined E_{BB} in this way, the mathematical equivalence in (1) is straightforward. But the WCP, if it is applied here, asserts *we ought not* to base inference on this statistic.

WCP is Accepted

The WCP says that if a mixture experiment outputs $(E'', .196)$, the inference should be calculated conditionally, that is, based on the sampling distribution of E''. Applying WCP we have the following:

The inference from $(E'', y''^* = .196)$ should be based on $\text{Infr}_{E''}(.196) = .37$. Likewise, if the mixture experiment resulted in $(E', .196)$, the inference should be based on the sampling distribution of E': $\text{Infr}_{E'}(.196) = .05$.

As was argued in Chapter 7(II), an argument for applying WCP is based on normative epistemological considerations of the way to make error probabilities relevant for specific inferences. This is why it may be open to dispute by those who hold, for a given context, that only average error rates matter in the long run. It is not about a mathematical identity, and in fact it is only because the unconditional analysis differs from the conditional one that it makes sense to argue that the former leads to counterintuitive inferences that are remedied by the distinct analysis offered by the latter.

So no contradiction results! Equivalently the preceding argument is invalid.

5 The Argument Formally Construed

Philosophers may wish to see the formal rendering of the argument.

Let A be as follows: outcome (E^j, y^{j*}) should be analyzed unconditionally as in Birnbaum experiment E_{BB}.

Then the WCP asserts not-A:

Not-A: outcome (E^j, y^{j*}) should be analyzed conditonally using the sampling distribution of E^j.

Premise 1: If A, then the inference from (E', y'^*) should be equal to the inference from (E'', y''). (The converse holds as well.)

Premise 2: Not-A.

Premise 2a: If not-A then the inference from (E'', y''^*) should be equal to the inference from y''^*.

Premise 2b: If not-A then the inference from (E', y'^*) should be equal to the inference from y'^*.

Therefore, the inference from y''^* should be equal to the inference from y'^*, the SLP.

From premises 2, 2a, and 2b, we have the following:

The inference from (E'', y''^*) should be equal to the inference from y''^*. The inference from (E', y'^*) should be equal to the inference from y'^*.

But nevertheless we can have that the inference from y''^* should NOT be equal to the inference from y'^*.

This demonstrates the argument is invalid.

The peculiarity of Birnbaum's experiment and its differences from the ordinary mixture have been duly noted (e.g., Cox and Hinkley, 1974, p. 41; Durbin, 1970), and many take it alone as grounds that Birnbaum's argument is not compelling. We see that an even stronger criticism is warranted.

My argument holds identically for derivation of the SLP based on a weaker condition than SP (see Evans et al., 1986), which I do not discuss here.

6 Concluding Remark

Chapter 7(II) grew out of numerous exchanges with David Cox in the 2 years since ERROR06. My goal was to extract and articulate the philosophical grounds underlying Cox's work on the nature and justification of types

of conditioning in relation to the goal of ensuring that error statistical assessments are relevant for the tasks of objective statistical inference. It became immediately clear to me that I would have to resolve the Birnbaum SLP puzzle if 7(II) was to consistently fill in the pieces of our earlier work in 7(I), hence, the birth of this essay. Fortunately, spotting the fallacy proved easy (in July 2007). To explain it as simply as I would have liked proved far more challenging.

References

Berger, J.O., and Wolpert, R.L. (1988), *The Likelihood Principle*, California Institute of Mathematical Statistics, Hayward, CA.

Birnbaum, A. (1962), "On the Foundations of Statistical Inference" (with discussion), *Journal of the American Statistical Association*, 57: 269–326.

Birnbaum, A. (1970), "More on Concepts of Statistical Evidence," *Journal of the American Statistical Association*, 67: 858–61.

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

Cox, D.R. (1977), "The Role of Significance Tests" (with discussion), *Scandinavian Journal of Statistics*, 4: 49–70.

Cox, D.R. and Hinkley, D.V. (1974), *Theoretical Statistics*, Chapman and Hall, London.

Durbin, J. (1970), "On Birnbaum's Theorem and the Relation between Sufficiency, Conditionality and Likelihood," *Journal of the American Statistical Association*, 65: 395–8.

Evans, M., Fraser, D.A.S., and Monette, G. (1986), "Likelihood," *Canadian Journal of Statistics*, 14: 180–90.

Lee, P.M. (2004), *Bayesian Statistics: an Introduction*, 3rd ed., Hodder Arnold, New York.

Lehmann, E.L. (1981), "An Interpretation of Completeness in Basu's Theorem," *Journal of the American Statistical Association*, 76: 335–40.

Mayo, D.G. and Cox, D.R. (2006), "Frequentist Statistics as a Theory of Inductive Inference," in J. Rojo (ed.), *Optimality: The Second Erich L. Lehmann Symposium*, Lecture Notes-Monograph Series, Institute of Mathematical Statistics (IMS), Vol. 49: 77–97.

Robins, J., and Wasserman, L. (2000), "Conditioning, Likelihood, and Coherence: A Review of Some Foundational Concepts," *Journal of the American Statistical Association*, 95: 1340–6.

Royall, R.M. (1997), *Statistical Evidence: A Likelihood Paradigm*, Chapman & Hall, London.

Savage, L., ed. (1962a), *The Foundations of Statistical Inference: A Discussion*. Methuen, London.

Savage, L. (1962b), "'Discussion on Birnbaum (1962)," *Journal of the American Statistical Association*, 57: 307–8.