Table of Contents: Supplemental Materials

Supplementary materials

James O. Berger, Robert L. Wolpert, M. J. Bayarri, M. H. DeGroot, Bruce M. Hill, David A. Lane and Lucien LeCam (1988) *The Likelihood Principle*, Lecture Notes-Monograph Series, Vol. 6, Institute of Mathematical Statistics. Stable URL: http://www.jstor.org/stable/4355509

pp. 6-7:

EXAMPLE 2. Suppose a substance to be analyzed can be sent either to a laboratory in New York or a laboratory in California. The two labs seem equally good, so a fair coin is flipped to choose between them, with "heads" denoting that the lab in New York will be chosen. The coin is flipped and comes up tails, so the California lab is used. After awhile, the experimental results come back and a conclusion must be reached. Should this conclusion take into account the fact that the coin eould have been heads, and hence that the experiment in New York might have been performed instead?

This, of course, is a variant of the famous Cox example (Cox (1958)- see also Cornfield (1969)), which concerns being given (at random) either an accurate or an inaccurate measuring instrument (and knowing which was given). Should the conclusion reached by experimentation depend only on the instrument actually used, or should it take into account that the other instrument might have been obtained?

In symbolic form, we can phrase this example as a "mixed experiment" in which with probabilities 1/2 (independent of θ) either experiment E_1, or experiment E_2 (both pertaining to ?) will be performed. Should the analysis depend only on the experiment actually performed, or should the possibility of having done the other experiment be taken into account?

The obvious intuitive answer to the questions in the above example is that only the experiment actually performed should matter. But this is counter to pre-experimental frequentist reasoning, which says that one should average over all possible outcomes (here, including the coin flip). One could argue that it is correct to condition on the coin flip, and then use the frequentist measures for the experiment actually performed, but the LP dis-allows this and is (surprisingly) derivable simply from conditioning on the coin flip plus sufficiency (see Chapter 3).

pp. 25-26:

The Conditionality Principle essentially says that, if an experiment is selected by some random mechanism independent of θ, then only the experiment actually performed is relevant. (The selection mechanism is ancillary, so this is a version of conditioning on an ancillary statistic.) The general conditionality principle is not needed here. Indeed we need only the following considerably weaker principle, named by Basu (1975).

WEAK CONDITIONALITY PRINCIPLE (WCP). Suppose there are two experiments $E_1 = (X_1, \theta, \{f_\theta^1\})$ and $E_2 = (X_2, \theta, \{f_\theta^2\})$, *where only the unknown parameter θ need be common to the two experiments. Consider the* <u>*mixed*</u> *experiment* E*, *whereby* J = 1 or 2 *is observed, each having*

probability ½ (independent of θ, X_1 or X_2) *and experiment* E_J, *is then performed. Formally,* $E^* = (X^*, \theta, \{f_\theta^*\})$, where $X^* = (J, X_J)$ *and* $f_\theta^*((j, x_j)) = \frac{1}{2} f_\theta^j(x_j)$. *Then,*

$$Ev(E^*, (j, x_j)) = Ev(E_j, x_j),$$

i.e., the evidence about θ f*rom* E^* *is just the evidence from the experiment actually performed.*

The WCP is nothing but a formalization of Example 2, and hence is essentially due to Cox (1958). It is hard to disbelieve the WCP, yet, as mentioned after Example 2, even the WCP alone has serious consequences.

FORMAL SUFFICIENCY PRINCIPLE: Consider experiment $E = (\mathbf{X}, \theta, \{f(\mathbf{x}|\theta)\})$ and suppose $T(\mathbf{X})$ is a sufficient statistic for θ. If $\mathbf{x}$ and $\mathbf{y}$ are sample points satisfying $T(\mathbf{x}) = T(\mathbf{y})$, then $\text{Ev}(E, \mathbf{x}) = \text{Ev}(E, \mathbf{y})$.

Thus, the *Formal* Sufficiency Principle goes slightly further than the Sufficiency Principle of Section 6.2. There no mention was made of the experiment. Here, we are agreeing to equate evidence if the sufficient statistics match. The Likelihood Principle can be derived from the Formal Sufficiency Principle and the following principle, an eminently reasonable one.

CONDITIONALITY PRINCIPLE: Suppose that $E_1 = (\mathbf{X}_1, \theta, \{f_1(\mathbf{x}_1|\theta)\})$ and $E_2 = (\mathbf{X}_2, \theta, \{f_2(\mathbf{x}_2|\theta)\})$ are two experiments, where only the unknown parameter θ need be common between the two experiments. Consider the mixed experiment in which the random variable J is observed, where $P(J = 1) = P(J = 2) = \frac{1}{2}$ (independent of θ, $\mathbf{X}_1$, or $\mathbf{X}_2$), and then experiment E_J is performed. Formally, the experiment performed is $E^* = (\mathbf{X}^*, \theta, \{f^*(\mathbf{x}^*|\theta)\})$, where $\mathbf{X}^* = (j, \mathbf{X}_j)$ and $f^*(\mathbf{x}^*|\theta) = f^*((j, \mathbf{x}_j)|\theta) = \frac{1}{2} f_j(\mathbf{x}_j|\theta)$. Then

$$\text{Ev}(E^*, (j, \mathbf{x}_j)) = \text{Ev}(E_j, \mathbf{x}_j). \tag{6.3.2}$$

The Conditionality Principle simply says that if one of two experiments is randomly chosen and the chosen experiment is done, yielding data $\mathbf{x}$, the information about θ *depends only on the experiment performed.* That is, it is the same information as would have been obtained if it were decided (nonrandomly) to do that experiment from the beginning, and data $\mathbf{x}$ had been observed. The fact that this experiment was performed, rather than some other, has not increased, decreased, or changed knowledge of θ.

Example 6.3.5 (Binomial/negative binomial experiment) Suppose the parameter of interest is the probability p, $0 < p < 1$, where p denotes the probability that a particular coin will land "heads" when it is flipped. Let E_1 be the experiment consisting of tossing the coin 20 times and recording the number of heads in those 20 tosses. E_1 is a binomial experiment and $\{f_1(x_1|p)\}$ is the family of binomial$(20, p)$ pmfs. Let E_2 be the experiment consisting of tossing the coin until the seventh head occurs and recording the number of tails before the seventh head. E_2 is a negative binomial experiment. Now suppose the experimenter uses a random number table to choose between these two experiments, happens to choose E_2, and collects data consisting of the seventh head occurring on trial 20. The Conditionality Principle says that the information about θ that the experimenter now has, $\text{Ev}(E^*, (2, 13))$, is the same as that which he would have, $\text{Ev}(E_2, 13)$, if he had just chosen to do the negative binomial experiment and had never contemplated the binomial experiment. ||

The following Formal Likelihood Principle can now be derived from the Formal Sufficiency Principle and the Conditionality Principle.

FORMAL LIKELIHOOD PRINCIPLE: Suppose that we have two experiments, $E_1 = (\mathbf{X}_1, \theta, \{f_1(\mathbf{x}_1|\theta)\})$ and $E_2 = (\mathbf{X}_2, \theta, \{f_2(\mathbf{x}_2|\theta)\})$, where the unknown parameter θ is the same in both experiments. Suppose $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ are sample points from E_1 and

E_2, respectively, such that

$$L(\theta|\mathbf{x}_2^*) = CL(\theta|\mathbf{x}_1^*) \tag{6.3.3}$$

for all θ and for some constant C that may depend on $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ but not θ. Then

$$\text{Ev}(E_1, \mathbf{x}_1^*) = \text{Ev}(E_2, \mathbf{x}_2^*).$$

The Formal Likelihood Principle is different from the Likelihood Principle in Section 6.3.1 because the Formal Likelihood Principle concerns two experiments, whereas the Likelihood Principle concerns only one. The Likelihood Principle, however, can be derived from the Formal Likelihood Principle by letting E_2 be an exact replicate of E_1. Thus, the two-experiment setting in the Formal Likelihood Principle is something of an artifact and the important consequence is the following corollary, whose proof is left as an exercise. (See Exercise 6.32.)

LIKELIHOOD PRINCIPLE COROLLARY: If $E = (\mathbf{X}, \theta, \{f(\mathbf{x}|\theta)\})$ is an experiment, then $\text{Ev}(E, \mathbf{x})$ should depend on E and $\mathbf{x}$ only through $L(\theta|\mathbf{x})$.

Now we state Birnbaum's Theorem and then investigate its somewhat surprising consequences.

Theorem 6.3.6 (Birnbaum's Theorem) *The Formal Likelihood Principle follows from the Formal Sufficiency Principle and the Conditionality Principle. The converse is also true.*

Proof: We only outline the proof, leaving details to Exercise 6.33. Let E_1, E_2, $\mathbf{x}_1^*$, and $\mathbf{x}_2^*$ be as defined in the Formal Likelihood Principle, and let E^* be the mixed experiment from the Conditionality Principle. On the sample space of E^* define the statistic

$$T(j, \mathbf{x}_j) = \begin{cases} (1, \mathbf{x}_1^*) & \text{if } j = 1 \text{ and } \mathbf{x}_1 = \mathbf{x}_1^* \text{ or if } j = 2 \text{ and } \mathbf{x}_2 = \mathbf{x}_2^* \\ (j, \mathbf{x}_j) & \text{otherwise.} \end{cases}$$

The Factorization Theorem can be used to prove that $T(J, \mathbf{X}_J)$ is a sufficient statistic in the E^* experiment. Then the Formal Sufficiency Principle implies

$$\text{Ev}(E^*, (1, \mathbf{x}_1^*)) = \text{Ev}(E^*, (2, \mathbf{x}_2^*)), \tag{6.3.4}$$

the Conditionality Principle implies

$$\begin{aligned} \text{Ev}(E^*, (1, \mathbf{x}_1^*)) &= \text{Ev}(E_1, \mathbf{x}_1^*) \\ \text{Ev}(E^*, (2, \mathbf{x}_2^*)) &= \text{Ev}(E_2, \mathbf{x}_2^*), \end{aligned} \tag{6.3.5}$$

and we can deduce that $\text{Ev}(E_1, \mathbf{x}_1^*) = \text{Ev}(E_2, \mathbf{x}_2^*)$, the Formal Likelihood Principle.

To prove the converse, first let one experiment be the E^* experiment and the other E_j. It can be shown that $\text{Ev}(E^*, (j, \mathbf{x}_j)) = \text{Ev}(E_j, \mathbf{x}_j)$, the Conditionality Principle. Then, if $T(\mathbf{X})$ is sufficient and $T(\mathbf{x}) = T(\mathbf{y})$, the likelihoods are proportional and the Formal Likelihood Principle implies that $\text{Ev}(E, \mathbf{x}) = \text{Ev}(E, \mathbf{y})$, the Formal Sufficiency Principle. □

Example 6.3.7 (Continuation of Example 6.3.5) Consider again the binomial and negative binomial experiments with the two sample points $x_1 = 7$ (7 out of 20 heads in the binomial experiment) and $x_2 = 13$ (the 7th head occurs on the 20th flip of the coin). The likelihood functions are

$$L(p|x_1 = 7) = \binom{20}{7} p^7(1-p)^{13} \quad \text{for the binomial experiment}$$

and

$$L(p|x_2 = 13) = \binom{19}{6} p^7(1-p)^{13} \quad \text{for the negative binomial experiment.}$$

These are proportional likelihood functions, so the Formal Likelihood Principle states that the same conclusion regarding p should be made in both cases. In particular, the Formal Likelihood Principle asserts that the fact that in the first case sampling ended because 20 trials were completed and in the second case sampling stopped because the 7th head was observed is immaterial as far as our conclusions about p are concerned. Lindley and Phillips (1976) give a thorough discussion of the binomial–negative binomial inference problem. ‖

This point, of equivalent inferences from different experiments, may be amplified by considering the sufficient statistic, T, defined in the proof of Birnbaum's Theorem and the sample points $\mathbf{x}_1^* = 7$ and $\mathbf{x}_2^* = 13$. For any sample points in the mixed experiment, other than $(1, 7)$ or $(2, 13)$, T tells which experiment, binomial or negative binomial, was performed and the result of the experiment. But for $(1, 7)$ and $(2, 13)$ we have $T(1, 7) = T(2, 13) = (1, 7)$. If we use only the sufficient statistic to make an inference and if $T = (1, 7)$, then all we know is that 7 out of 20 heads were observed. We do not know whether the 7 or the 20 was the fixed quantity.

Many common statistical procedures violate the Formal Likelihood Principle. With these procedures, different conclusions would be reached for the two experiments discussed in Example 6.3.5. This violation of the Formal Likelihood Principle may seem strange because, by Birnbaum's Theorem, we are then violating either the Sufficiency Principle or the Conditionality Principle. Let us examine these two principles more closely.

The Formal Sufficiency Principle is, in essence, the same as that discussed in Section 6.1. There, we saw that all the information about θ is contained in the sufficient statistic, and knowledge of the entire sample cannot add any information. Thus, basing evidence on the sufficient statistic is an eminently plausible principle. One shortcoming of this principle, one that invites violation, is that it is very model-dependent. As mentioned in the discussion after Example 6.2.9, belief in this principle necessitates belief in the model, something that may not be easy to do.

Most data analysts perform some sort of "model checking" when analyzing a set of data. Most model checking is, necessarily, based on statistics other than a sufficient statistic. For example, it is common practice to examine *residuals* from a model, statistics that measure variation in the data not accounted for by the model. (We will see residuals in more detail in Chapters 11 and 12.) Such a practice immediately violates the Sufficiency Principle, since the residuals are not based on sufficient statistics.

(Of course, such a practice directly violates the Likelihood Principle also.) Thus, it must be realized that *before* considering the Sufficiency Principle (or the Likelihood Principle), we must be comfortable with the model.

The Conditionality Principle, stated informally, says that "only the experiment actually performed matters." That is, in Example 6.3.5, if we did the binomial experiment, and not the negative binomial experiment, then the (not done) negative binomial experiment should in no way influence our conclusion about θ. This principle, also, seems to be eminently plausible.

How, then, can statistical practice violate the Formal Likelihood Principle, when it would mean violating either the Principle of Sufficiency or Conditionality? Several authors have addressed this question, among them Durbin (1970) and Kalbfleisch (1975). One argument, put forth by Kalbfleisch, is that the proof of the Formal Likelihood Principle is not compelling. This is because the Sufficiency Principle is applied in ignorance of the Conditionality Principle. The sufficient statistic, $T(J, \mathbf{X}_J)$, used in the proof of Theorem 6.3.6 is defined on the mixture experiment. If the Conditionality Principle were invoked first, then separate sufficient statistics would have to be defined for each experiment. In this case, the Formal Likelihood Principle would no longer follow. (A key argument in the proof of Birnbaum's Theorem is that $T(J, \mathbf{X}_J)$ *can take on the same value for sample points from each experiment.* This cannot happen with separate sufficient statistics.)

At any rate, since many intuitively appealing inference procedures do violate the Likelihood Principle, it is not universally accepted by all statisticians. Yet it is mathematically appealing and does suggest a useful data reduction technique.

6.4 The Equivariance Principle

The previous two sections both describe data reduction principles in the following way. A function $T(\mathbf{x})$ of the sample is specified, and the principle states that if $\mathbf{x}$ and $\mathbf{y}$ are two sample points with $T(\mathbf{x}) = T(\mathbf{y})$, then the same inference about θ should be made whether $\mathbf{x}$ or $\mathbf{y}$ is observed. The function $T(\mathbf{x})$ is a sufficient statistic when the Sufficiency Principle is used. The "value" of $T(\mathbf{x})$ is the set of all likelihood functions proportional to $L(\theta|\mathbf{x})$ if the Likelihood Principle is used. The Equivariance Principle describes a data reduction technique in a slightly different way. In any application of the Equivariance Principle, a function $T(\mathbf{x})$ is specified, but if $T(\mathbf{x}) = T(\mathbf{y})$, then the Equivariance Principle states that the inference made if $\mathbf{x}$ is observed should have a *certain relationship* to the inference made if $\mathbf{y}$ is observed, although the two inferences may not be the same. This restriction on the inference procedure sometimes leads to a simpler analysis, just as do the data reduction principles discussed in earlier sections.[2]

Although commonly combined into what is called the Equivariance Principle, the data reduction technique we will now describe actually combines two different equivariance considerations.

[2] As in many other texts (Schervish 1995; Lehmann and Casella 1998; Stuart, Ord, and Arnold 1999) we distinguish between *equivariance*, in which the estimate changes in a prescribed way as the data are transformed, and *invariance*, in which the estimate remains unchanged as the data are transformed.

We may abbreviate this as follows:

If $\boldsymbol{s}$ is minimal sufficient for $\boldsymbol{\theta}$ in experiment E, and $\boldsymbol{s}(\boldsymbol{y}') = \boldsymbol{s}(\boldsymbol{y}'')$, then the inference from $\boldsymbol{y}'$ and $\boldsymbol{y}''$ about θ should be identical; that is, $\text{Infr}_E(\boldsymbol{y}') = \text{Infr}_E(\boldsymbol{y}'')$.

However, when proposing to apply the sufficiency principle to a particular inference account, the relevant method for inference must be taken into account. That is, Infr_E is relative to the inference account.

7.2 Sufficiency in Sampling Theory

If a random variable $\boldsymbol{Y}$, in a given experiment E, arises from $f(\boldsymbol{y}; \theta)$, and the assumptions of the model are valid, then all the information about θ contained in the data is obtained from consideration of its minimal sufficient statistic $\boldsymbol{S}$ and its *sampling distribution* $f_S(\boldsymbol{s}; \theta)$.

An inference in sampling theory, therefore, needs to include the relevant sampling distribution, whether it was for testing or estimation. Thus, in using the abbreviation $\text{Infr}_E(\boldsymbol{y})$ to refer to an inference from $\boldsymbol{y}$ in a sampling theory experiment E, we assume for simplicity that E includes a statement of the probability model, parameters, and sampling distribution corresponding to the inference in question. This abbreviation emphasizes that the inference that is licensed is relative to the particular experiment, the type of inference, and the overall statistical approach being discussed.[2]

In the case of frequentist sampling theory, features of the experiment that alter the sampling distribution must be taken account of in determining what inferences about θ are warranted, and when the same inferences from given experiments may be drawn. Even if $\boldsymbol{y}'$ and $\boldsymbol{y}''$ have proportional likelihoods but are associated with different relevant sampling distributions, corresponding to E' and E'', $\boldsymbol{y}'$ and $\boldsymbol{y}''$ each provides different relevant information for inference. It is thus incorrect to suppose, within the sampling paradigm, that it is appropriate to equate $\text{Infr}_{E'}(\mathbf{y}')$ and $\text{Infr}_{E''}(\mathbf{y}'')$.

These points show that sampling theory violates what is called the *strong likelihood principle.*

[2] This abbreviation, like Birnbaum's Ev(E, **x**), may be used to discuss general claims about principles of evidence. Birnbaum's Ev(E, **x**), "the evidence about the parameter arising from experiment E and result **x**," is, for Birnbaum, the inference, conclusion or report, and thus is in sync with our notion (Birnbaum, 1962).

We prefer it because it helps avoid assuming a single measure of "the" evidence associated with an experimental outcome. By referring to the inference licensed by the result, it underscores the need to consider the associated methodology and context.

7.3 The Strong Likelihood Principle (SLP)

Suppose that we have *two* experiments, E' and E'', with different probability models $f'_{Y'}(y';\theta)$ and $f''_{Y''}(y'';\theta)$, 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 $\mathbf{y}'^*$ and $\mathbf{y}''^*$ are proportional, *then* $\mathbf{y}'^*$ and $\mathbf{y}''^*$ have the identical evidential import for any inference about θ.

Here proportionality means that, for all θ, $f''_Y(\boldsymbol{y}'';\theta)/f'_Y(\boldsymbol{y}';\theta)$ is equal to a constant that does not depend on θ. A sample of, say, six successes in twenty trials would, according to the SLP, have the identical evidential import whether it came from a binomial experiment, with sample size fixed at twenty, or from a negative binomial experiment where it took twenty trials to obtain six successes.

By contrast, suppose a frequentist is interested in making an inference about θ on the basis of data $\boldsymbol{y}'$ consisting of r successes in n trials in a binomial experiment E'. Relevant information would be lost if the report were reduced to the following: there were r successes in n Bernoulli trials, generated from *either* a binomial experiment with n fixed, $\boldsymbol{y}'^*$, or a negative binomial experiment with r fixed, $\boldsymbol{y}''^*$ – concealing which was actually the source of the data. Information is lost because $\text{Infr}_{E'}(\boldsymbol{y}'^*)$ is *not* equal to $\text{Infr}_{E''}(\boldsymbol{y}''^*)$ due to the difference in the associated sampling distributions. Equivalences that hold with respect to a single experiment, as is the case with sufficiency, cannot be assumed to hold in comparing data from different experiments.

8 Sufficient Statistics and Test Statistics

How then are we to extract answers to the research question out of $f_S(\boldsymbol{s};\boldsymbol{\theta})$; all that the reduction to $\boldsymbol{s}$ has done is to reduce the dimensionality of the data. To establish a significance test, we need to choose an appropriate test statistic $T(\boldsymbol{Y})$ and find a distribution for assessing its concordancy with H_0. To warrant the interpretations of the various significance tests that we delineated in the first part of this chapter (Mayo and Cox), we need to consider how to identify test statistics to construct appropriate tests.

To interpret t, the observed value of T, we compare it with its predicted value under the null hypothesis by finding, for any observed value t, $p = P(T \geq t; H_0)$. That is, we examine how extreme t is in its probability distribution under H_0. Thus, we need both to choose an appropriate test statistic $T(\boldsymbol{Y})$ and also to compute its distribution in order to compare t with what is expected under H_0. To this end we find a suitable feature t of the data,

explicit form in that context, would reject the Probability Mixture Assumption, primarily on psychological grounds.

In the present context, the difficulties are not so much psychological as measure-theoretical. Is the probability measure on the ancillary statistic (denoted as h by Professor Birnbaum, equivalent to p in my examples) homogeneous in measure with the probability measure on the experimental outcome x_h—is it, that is to say, of the same evidential *quality* for purposes of inference?

In conclusion, I would like to urge Professor Birnbaum, before he goes on to erect an elaborate superstructure on these undoubtedly very impressive "Foundations," to analyze in more detail the implications of his assumptions. Research in this area could be of great value to all of us.

A. P. DEMPSTER: My comments stem from the uneasiness which I habitually feel when confronted with strong, apparently binding consequences of a few simple and plausible postulates. I tend to think that the world really cannot be so uncomplicated.

In this instance I am asked to accept the *likelihood principle* (L) because it follows from the *conditionality principle* (C) together with the *sufficiency principle* (S). But are there not other assumptions and principles which, although stated less explicitly, play important roles in the argument leading to (L)? I would like to single out such a principle, to be called here the *uniqueness principle* (U). This may be stated as follows.

The uniqueness principle (U): For a given statistician with a given mathematical-statistical model and a given set of observations, there is one and only one evidential meaning.

It would be possible to introduce a stronger version of (U), say (U'), which would require all statisticians to agree on evidential meaning. I do not know whether Professor Birnbaum has (U) or (U') in mind. It is clear, however, that at least (U) is firmly implicated in the argument of Part I of the paper.

My two chief comments are as follows. Firstly, I would like to suggest that there is no clear positive rationale for accepting (U). Could it not be, for example, that the model and data leave undetermined some ingredient essential to evidential meaning? Or, perhaps, could it not be that the concept of a unique evidential meaning is simply an unrealistic ideal? Secondly, I would like to suggest that leaning too heavily on (U) can only lead to extreme, and therefore unconvincing, standpoints on philosophical statistics. For example, could many statisticians really accept the likelihood principle? (L) implies exclusion from any role in evidential meaning of significance tests, confidence statements, and even so basic a concept as the mean square error of an estimator. To eradicate such concepts from the thought processes of statisticians would require a prodigious brain-washing program. Would such a program really rest on firm philosophical grounds? I, at least, would prefer to relax the uniqueness principle somewhat.

An incidental comment is that Professor Birnbaum appears to relax (U) in Part II of his paper when he allows his "intrinsic" methods to refer to "conventional experimental frames of reference." The key word is "conventional." Surely this introduces a logical weakness into the paper. For, if (U) can be

given up so easily in Part II, then how can I find Part I convincing when it relies on (U)? On the other hand, if intrinsic methods are left outside of the canon, then direct non-Bayesian interpretation of likelihoods lacks an intuitive basis, at least to me.

OSCAR KEMPTHORNE:* I was a bit surprised at the emphasis on "experiment" "experimental situation" etc. in Dr. Birnbaum's paper. It is a truism, I believe, that there is never an adequate mathematical statistical model for any actual situation. The "reporting" of "experimental results in journals" of the empirical sciences must of necessity be incomplete not only with regard to the condensation of the data, and the drawing of conclusions from them, but also in the description of the performance of the experiment. And the interpretation of experimental results depends as much on the faith put by the interpreter in the completeness and accuracy of the description of what was done as in what the data indicate.

I found, however, much of interest and stimulation in Dr. Birnbaum's paper. I take the problem to be the purely logical one that the observation x has a probability $f(x, \theta)$ and one wishes to characterize the evidence about θ supplied by x and the postulation of the function $f(x, \theta)$.

I learned much of what I know of statistics by reading Fisher's papers and it is, I believe, implicit in these dating back at least to 1922 (Mathematical Foundations of Theoretical Statistics) that the probability of the observed sample is the basis for any feeling we may have about the parameter's values. It therefore does not strike me as appropriate to cite Fisher 1956 as the reference for Fisher's views on likelihood except that it is evidence that Fisher has not changed his mind over a period of some 30 years or so. It seems obvious to me that the probability (or probability density) with other information such as the origin of the data is all we have, and that the idea of sufficiency arose solely out of considering the probability or the likelihood. I was therefore surprised to see that Dr. Birnbaum finds S appropriate, and makes it his starting point, when one can arrive at a sufficient statistic only by considering the probability (density) of the sample. It may be that I am quite out of touch with what mathematicians have done with *Fisher's* concept of sufficiency. The fact that one does not wish to modify a binomial observation by introducing a random variable to smooth out the probability is not to me an argument for considering the sufficient statistic but an argument for not introducing an irrelevancy; or to put things a different way, for not saying I observed $5+r$ successes in 10 trials when I observed 5 successes.

I found Dr. Birnbaum's paper somewhat obscure as regards the relationships among S, C, and L. He states

(1) "S is implied mathematically by C"
(2) Lemma 1: L implies S
(3) Lemma 2: L implies and is implied by S and C.

In particular, because C implies S, lemma 2 should read in part, it seems, C implies L. From the point of view of presentation, one can use $L \rightarrow C$ and $C \rightarrow S$

* Mr. Kempthorne was unable to attend. His written comments were communicated to the author and the editor after the meeting.

drawn about θ cannot depend on the particular sampling scheme adopted.

These results are very special cases of ones applying whenever we have a "stopping rule" depending in some way on the data currently accumulated but not on further information about the unknown parameter.

Example 2.34. Sequential sampling. Suppose that observations are taken one at a time and that after each observation a decision is taken as to whether to take one more observation. Given $m-1$ observations $y_1, \ldots, y_{m-1}$, there is a probability $p_{m-1}(y_1, \ldots, y_{m-1})$ that one more observation is in fact taken. The conditional p.d.f. of Y_m given $Y_1 = y_1, \ldots, Y_{m-1} = y_{m-1}$ is written in the usual way. Note that this includes very general forms of sequential sampling in which observations may be taken singly or in groups.

Suppose that the data are $(n, y_1, \ldots, y_n)$. Then the likelihood, i.e. the joint probability that observations are taken in the way specified and give the values actually observed, is

$$p_0 f_{Y_1}(y_1; \theta) p_1(y_1) f_{Y_2|Y_1}(y_2|y_1; \theta) \ldots p_{n-1}(y_1, \ldots, y_{n-1})$$

$$f_{Y_n|Y_{n-1},\ldots,Y_1}(y_n|y_{n-1}, \ldots, y_1; \theta)\{1 - p_n(y_1, \ldots, y_n)\}.$$

Thus, so long as the probabilities defining the sampling scheme are known they form a constant factor in the likelihood function and the dependence on the parameters is fixed by the observations actually obtained, in fact by the joint p.d.f. of $Y_1, \ldots, Y_n$. Therefore, if the strong likelihood principle were accepted, the conclusion to be drawn about θ would be the same as if n were fixed. Note, however, that N is not in general an ancillary statistic and that conditioning on its value is not a consequence of the conditionality principle as formulated above.

We noted at the end of the previous subsection that the weak likelihood principle and the sufficiency principle are equivalent. The deduction of the strong likelihood principle from the sufficiency principle plus some form of the conditionality principle has been considered by Birnbaum (1962, 1969, 1970), Barnard, Jenkins and Winsten (1962), Durbin (1970), Savage (1970), Kalbfleisch (1974) and Basu (1973). We shall not go into details, but the following seems the essence of the matter.

Suppose that (44) holds. Now pretend that we have the following experiment, which we call the enlarged experiment:

observe Y with probability $\frac{1}{2}$,

or

observe Z with probability $\frac{1}{2}$.

Now imagine this experiment done and consider the following two outcomes:

(a) Y is observed and $Y = y$,

(b) Z is observed and $Z = z$.

Now the likelihood functions of (a) and (b) in the enlarged experiment being $\frac{1}{2}f_Y(y\,;\theta)$ and $\frac{1}{2}f_Z(z\,;\theta)$, they are by (44) proportional for all θ. Hence, if we can apply the weak likelihood principle or sufficiency principle to the enlarged experiment, then the conclusions from (a) and (b) should be identical. Finally in, say, (a) the event "Y is observed" has fixed probability $\frac{1}{2}$ (c.f. Example 2.26) and if the conditionality principle is applied to the enlarged experiment the inference from (a) should be the same as from the simpler component experiment in which $Y = y$. Similarly for (b), and so the strong likelihood principle has been deduced.

There are several reasons why this argument is not compelling. One (Durbin, 1970) is that in (a) the event "Y is observed" will not be part of the minimal sufficient statistic and hence, at least on the definition used here, does not qualify to be an ancillary statistic. A second and more basic reason concerns the propriety of regarding the enlarged experiment as fit for the application of the weak likelihood principle. If we were to insist that calculations are made conditionally on the experiment actually performed, i.e. were to apply some form of conditionality principle before applying sufficiency, the basis for considering the enlarged experiment would collapse.

The Bayesian methods to be considered in Chapter 10 do satisfy the strong likelihood principle; nearly all the other methods do not.

(vi) Invariance principle

The very simplest form of invariance argument holds when there are two values y' and y'' of the vector random variable Y that have the same value of the p.d.f. for all θ, i.e. $f_Y(y'\,;\theta) = f_Y(y''\,;\theta)$ for all $\theta \in \Omega$. It is then, under the model, a pure convention which is called y' and which y'', i.e. we can interchange y' and y'' without material change of the situation. The invariance principle in this case requires

"something much more substantial." De Finetti called this "the involuntarily destructive aspect of Wald's work" (1972, p. 176). Cox remarks:

> [T]here is a distinction between the Neyman–Pearson formulation of testing regarded as clarifying the meaning of statistical significance via hypothetical repetitions and that same theory regarded as in effect an instruction on how to implement the ideas by choosing a suitable α in advance and reaching different decisions accordingly. The interpretation to be attached to accepting or rejecting a hypothesis is strongly context-dependent . . . (Cox 2006a, p. 36)

If N-P long-run performance concepts serve to clarify the meaning of statistical significance tests, yet are not to be applied literally, but rather in some inferential manner – call this the *meaning* vs. *application distinction* – the question remains – how?

My answer, in terms of severity, may be used whether you prefer the N-P tribe (tests or confidence intervals) or the Fisherian tribe. What would that most eminent Fisherian, Sir David Cox, say? In 2004, in a session we were in on statistical philosophy, at the semi-annual Lehmann conference, we asked: Was it possible to view "Frequentist Statistics as a Theory of Inductive Inference"? If this sounds familiar it's because it echoes a section from Neyman's quarrel with Carnap (Section 2.7), but how does a Fisherian answer it? We began "with the core elements of significance testing in a version very strongly related to but in some respects different from both Fisherian and Neyman–Pearson approaches, at least as usually formulated" (Mayo and Cox 2006, p. 80). First, there is no suggestion that the significance test would typically be the only analysis reported. Further, we agree that "the justification for tests will not be limited to appeals to long-run behavior but will instead identify an inferential or evidential rationale" (ibid., p. 81).

With N-P results available, it became easier to understand why intuitively useful tests worked for Fisher. N-P and Fisherian tests, while starting from different places, "lead to the same destination" (with few exceptions) (Cox 2006a, p. 25). Fisher begins with seeking a test statistic that reduces the data as much as possible, and this leads him to a *sufficient* statistic. Let's take a side tour to sufficiency.

Exhibit (ii): Side Tour of Sufficient Statistic. Consider n independent trials $\boldsymbol{X} := (X_1, X_2, \ldots, X_n)$ each with a binary outcome (0 or 1), where the probability of success is an unknown constant θ associated with Bernoulli trials. The number of successes in n trials, $Y = X_1 + X_2 + \cdots + X_n$ is Binomially distributed with parameters θ and n. The sample mean, which is just $\overline{X} = Y/n$, is a natural estimator of θ with a highly desirable property: it is *sufficient*, i.e., it is

a function of the *sufficient* statistic Y. Intuitively, a sufficient statistic reduces the n-dimensional sample $\boldsymbol{X}$ into a statistic of much smaller dimensionality without losing any relevant information for inference purposes. Y reduces the n-fold outcome $\boldsymbol{x}$ to one dimension: the number of successes in n trials. The parameter of the Binomial model θ also has one dimension (the probability of success on each trial).

Formally, a statistic Y is said to be sufficient for θ when the distribution of the sample is no longer a function of θ when conditioned on Y, i.e., $f(x \mid y)$ does not depend on θ,

$$f(x; \theta) = f(y; \theta)\, f(x|y).$$

Knowing the distribution of the sufficient statistic Y suffices to compute the probability of any data set $\boldsymbol{x}$. The test statistic $d(\boldsymbol{X})$ in the Binomial case is $\sqrt{n}(\overline{X} - \theta_0)/\sigma$, $\sigma = \sqrt{[\theta(1-\theta)]}$ and, as required, gets larger as $\overline{X}$ deviates from θ_0. Thanks to $\overline{X}$ being a function of the sufficient statistic Y, it is the basis for a test statistic with maximal sensitivity to inconsistencies with the null hypothesis.

The Binomial experiment is equivalent to having been given the data $\boldsymbol{x}_0 = (x_1, x_2, \ldots, x_n)$ in two stages (Cox and Mayo 2010, p. 285):

First, you're told the value of Y, the number of successes out of n Bernoulli trials. Then an inference can be drawn about θ using the sampling distribution of Y.

Second, you learn the value of the specific data, e.g., the first k trials are successes, the rest failure. The second stage is equivalent to observing a realization of the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Y} = \boldsymbol{y}$. If the model is appropriate then "the second phase is equivalent to a random draw from a totally known distribution." All permutations of the sequence of successes and failures are equally probable (ibid., pp 284–5).

"Because this conditional distribution is totally known, it can be used to assess the validity of the assumed model." (ibid.) Notice that for a given $\boldsymbol{x}$ *within* a given Binomial experiment, the ratio of likelihoods at two different values of θ depends on the data only through Y. This is called the *weak likelihood principle* in contrast to the general (or strong) LP in Section 1.5.

Principle of Frequentist Evidence, FEV

Returning to our topic, "Frequentist Statistics as a Theory of Inductive Inference," let me weave together three threads: (1) the Frequentist Principle of Evidence (Mayo and Cox 2006), (2) the divergent interpretations growing out of Cox's taxonomy of test hypotheses, and (3) the links to statistical