The Frequentist Viewpoint and Conditioning

by

James Berger*
Purdue University

Technical Report #83-48

Department of Statistics
Purdue University
West Lafayette, IN 47907

Revised, February 1984

*Research supported by the National Science Foundation under Grant MCS 81-01670.

Abstract

The relationship of the frequentist viewpoint and the conditional viewpoint in statistics is discussed. After an attempt to carefully define the frequentist viewpoint, it is shown that Kiefer's conditional and estimated confidence theories, particularly the latter, are very successful in dealing with many of the conditional difficulties of standard frequentist theory, while maintaining frequentist validity. Not all conflicts with the conditional viewpoint can be avoided, however, as shown by the likelihood principle and the stopping rule principle. These principles and the conflicts are briefly reviewed. Also, a justification for the stopping rule principle is given in terms of frequentist admissibility.

1. Introduction

This paper is based on a talk given in the memorial session for Jack Kiefer at the Berkeley Conference in honor of Kiefer and Neyman. The bulk of the talk was devoted to a review of Kiefer's work on conditional confidence and estimated confidence. The talk also touched on broader issues, such as what it means to be a frequentist, and what the basic issues in conditioning are. This paper will mainly be devoted to a discussion of these broader issues, partly because a review of Kiefer's work on the subject will appear in the volume of his collected works, partly because Kiefer's work can best be appreciated in a general discussion of the issues, and partly because the broader issues themselves deserve more exposure than they are commonly accorded.

Section 2 of the paper begins with a discussion of what it means to be a frequentist, based on the original views of the developer of the frequentist school, Jerzy Neyman. This issue deserves attention because Neyman's original justification for the frequentist position (which strikes us as the best justification that has been given) is not the justification most commonly taught. Furthermore, the issue is important in determining the type of conditional frequentist theory that is most justifiable. Kiefer and Brownie discussed, in a series of papers (Kiefer (1975, 1976, 1977) and Brownie and Kiefer (1977)), two possible approaches to such a conditional frequentist theory, the "conditional confidence" and "estimated confidence" approaches. These are briefly reviewed in Section 2.2,followed, in Section 2.3, by a discussion of the very interesting problem of selection of a conditional frequentist procedure.

Section 3 turns to a discussion of conditioning from a more philosophical viewpoint, concentrating on such issues as the likelihood principle and the stopping rule principle. Formal developments are eschewed, in preference for more

intuitive presentation of the issues and discussion of conflicts and paradoxes. For instance, a not widely known paradox between the frequentist notion of admissibility and the idea that decisions should depend on the stopping rule is discussed. Section 4 presents some conclusions.

This paper is, in no sense, meant to be a thorough review of any of the topics discussed. Instead it is intended to serve as an introduction to a number of interesting, and too often ignored, issues. Extensive references are not given, and indeed no attempt has been made to trace back ideas to sources. More scholarly reviews of, and references for, these topics can be found in (for example) Kiefer (1977), Berger and Wolpert (1984), and Berger (1984a).

Before proceeding, some notation will be introduced. Also, a series of simple examples will be presented to give a feeling for the conditioning issues and to provide a background for later discussion.

It will be assumed that an "experiment" E is performed, which consists of observing a random quantity X having distribution P_θ on a sample space $\mathcal{X}$, $\theta \in \Theta$ being unknown. (For the most part θ will be taken to be a parameter, so that $\{P_\theta\}$ is a parametric family, but this need not necessarily be the case; θ could just index some nonparametric family.) When $\{P_\theta\}$ is a dominated family with respect to a measure ν, we will denote the density of X by

$$(1.1)\quad f(x|\theta) = d\,P_\theta(x)/d\nu(x) \quad .$$

Expectation over the distribution P_θ will be denoted by E_θ. Finally, the actual data from the experiment (i.e., the realization of X) will be denoted by x.

The following examples are well known illustrations of the conditioning problem. For historical references and other examples (many, of substantial practical importance) see Kiefer (1977), Berger (1984a), and Berger and Wolpert (1984).

Example 1. Suppose $X = (X_1, X_2)$, where X_1 and X_2 are independently distributed according to the distribution

$$P_\theta(X_i = \theta-1) = P_\theta(X_i = \theta+1) = \frac{1}{2},$$

where $-\infty<\theta<\infty$. Consider the "confidence procedure" defined by

$$C(x) = \begin{cases} \text{the point } \frac{1}{2}(x_1+x_2) & \text{if } |x_1-x_2| = 2 \\ \text{the point } x_1-1 & \text{if } |x_1-x_2| = 0. \end{cases}$$

Since (by an easy calculation)

$$P_\theta(C(X) \text{ contains } \theta) = .75 \text{ for all } \theta,$$

the confidence procedure C is a valid 75% frequentist confidence procedure, and furthermore satisfies any number of frequentist optimality properties. It is clearly misleading, however, to present C(x) and state "75% confidence" after seeing the data x, since if $|x_1-x_2| = 2$ one is absolutely certain that $\theta \in C(x)$, while if $|x_1-x_2| = 0$ one is (more or less) equally uncertain as to whether θ is x_1-1 or x_1+1. Conditional on the data x, one should state either 100% or 50% "confidence," depending on the value of $|x_1-x_2|$.

Example 2a. Suppose X is 1, 2, or 3 and θ is 1 or 2, with $P_\theta(x)$ given in the following table:

	X		
	1	2	3
P_0	.009	.001	.99
P_1	.001	.989	.01

The test, which accepts P_0 when x = 3 and accepts P_1 otherwise, is a most powerful test with both error probabilities equal to .01. Hence, it would be valid to make the frequentist statement, upon observing x = 1, "My test has rejected P_0 and the error probability is .01." Again this seems misleading, since the likelihood ratio is actually 9 to 1 in favor of P_0, which is being rejected.

Example 2b. One could object in Example 2a, that the .01 level test is inappropriate, and that one should use the .001 level test, which rejects only when x = 2. Consider, however, the following slightly changed version:

	X		
	1	2	3
P_0	.005	.005	.99
P_1	.0051	.9849	.01

.

Again the test which rejects P_0 when x = 1 or 2 and accepts otherwise has error probabilities equal to .01, and now it indeed seems sensible to take the indicated actions. (Suppose an action must be taken.) It still seems unreasonable, however, to report an error probability of .01 upon rejecting P_0 when x = 1, since the data provides very little evidence in favor of P_1.

Example 3(a). Suppose X is $\mathcal{N}(\theta,1)$, and that it is desired to test

$$H_0: \theta \leq -2 \text{ versus } H_a: \theta \geq 2.$$

Consider the test: reject H_0 if $x \geq 0$. Clearly, for $\theta < 0$,

$$P_\theta(\text{Type I error}) = P_{-\theta}(\text{Type II error}) \leq P_{-2}(X \geq 0) = .0228.$$

If x = 0 is observed, however, it seems misleading to state that "H_0 is rejected, and the error probability is at most .0228."

Example 3(b). Suppose X is $\mathcal{N}(\theta,1)$, and that it is desired to test

$$H_0: \theta \leq 0 \text{ versus } H_a: \theta > 0.$$

Consider the test: reject H_0 if $x \geq 0$. Clearly

$$P_{\theta=0} \text{ (Type I error)} = \frac{1}{2} .$$

If $x = 10$ is observed, we are virtually certain that H_a is true, and yet (via formal frequentist theory) all we can say is that we reject with error probability of $\frac{1}{2}$.

2. The Frequentist Approach to Conditioning

Before discussing how frequentists can deal with (at least many) conditioning problems, it is necessary to define what a valid frequentist approach is. Any such attempt is fraught with peril, since a large number of statsticians with quite conflicting beliefs call themselves "frequentists." What follows is a quite restrictive definition of a frequentist, which, however, hopefully contains the essence of what Neyman felt was the frequentist rationale. Some comments about other "frequentist" views will be given at the end of Section 2.1.

2.1 Frequentist Rationale

Until recently, I thought a frequentist was someone with the following approach:

(i) Select a procedure $\delta(x)$ for use;

(ii) Define a criterion (or loss) $L(\theta,\delta)$ that one would like to know or that measures performance;

(iii) Report $(\delta, R_\delta(\theta))$, where

$$R_\delta(\theta) = E_\theta L(\theta,\delta(X)).$$

Example 4. For confidence set problems, $\delta(x) = C(x) \subset \Theta$ defines a confidence procedure,

$$L(\theta, C(x)) = 1 - I_{C(x)}(\theta)$$

($I_A(\theta)$ denoting the usual indicator function) is what one would like to know and is the usual measure of performance, and

(2.1) $R_C(\theta) = E_\theta L(\theta, C(X)) = 1 - P_\theta(C(X) \text{ contains } \theta)$.

Example 5. In testing, let δ denote a test and L be zero-one loss, so that

$$R_\delta(\theta) = P_\theta(\text{incorrect decision}).$$

Of course, decision-theoretic examples are plentiful.

The usual justification for the frequentist viewpoint outlined above is that if, for a given θ_0, one were to repeatedly use δ_0 on (independent) $X_i \sim P_{\theta_0}$, then (with probability one under mild conditions)

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} L(\theta_0, \delta(X_i)) = R_\delta(\theta_0). \tag{2.2}$$

Thus, by reporting $R_\delta(\theta)$, one is giving the long run performance of δ for θ.

This leads, however, to two immediate questions:

(i) One does not know θ, so how is $R_\delta(\theta)$ to be used?

(ii) In reality one will be using δ with different problems having different θ_i, so what is the value in knowing (2.2)?

These questions are perplexing, and in an effort to understand the frequentist answers to them I returned to Neyman's original papers which developed the frequentist method. (See A Selection of Early Statistical Papers of J. Neyman; also relevant is Neyman (1957).) To my surprise, a different and more appealing frequentist viewpoint emerged from these papers, a viewpoint based on use of δ in different problems with different θ_i. In the notation of this paper, this viewpoint can be expressed as follows.

First, consider an infinite sequence of problems in which $X_i \sim P_{\theta_i}$ will be observed. Let $\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots)$. Suppose certain subsequences $\underset{\sim}{\theta}_\omega$, $\omega = (\omega(1), \omega(2)\ldots) \in \{1,2,\ldots\}^\infty$, are of interest.

Definition 1. A quantity $\overline{R}_\delta^\omega$ will be called a valid frequentist measure of the performance of δ on $\underset{\sim}{\theta}_\omega$ if, with probability one,

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} L(\theta_{\omega(i)}, \delta(X_{\omega(i)})) \leq \overline{R}_\delta^\omega . \tag{2.3}$$

The idea here is that one will report $(\delta, \overline{R}_\delta^\omega)$, and can be assured that, in repeated use of δ for $\theta_i \in \underset{\sim}{\theta}_\omega$, the average long run performance will be <u>at least</u> $\overline{R}_\delta^\omega$.

Often, as in typical estimation or confidence set problems, one will be interested only in $\omega = (1,2,\ldots)$, i.e., will want to find a quantity $\overline{R}_\delta$ such that, with probability one for all $\underset{\sim}{\theta}$,

$$(2.4) \quad \lim_{n\to\infty} \sup \frac{1}{n} \sum_{i=1}^{n} L(\theta_i, \delta(X_i)) \leq \overline{R}_\delta \ .$$

Indeed, it can be argued that this is the only situation in which a truly meaningful frequentist statement is being made, in that, without knowledge of the θ_i (or ω) that are likely to be encountered, a reporting of $\overline{R}_\delta^\omega$ does not completely convey the performance that is to be expected.

The typical situations in which subsequences are of interest are testing problems, where $\underset{\sim}{\theta}_{\omega_0}$ and $\underset{\sim}{\theta}_{\omega_1}$ may refer, say, to sequences of null and alternative hypotheses, respectively. Then, for zero-one loss, $\overline{R}_\delta^{\omega_0}$ and $\overline{R}_\delta^{\omega_1}$ would be bounds on the propabilities of Type I and Type II error, respectively. Of course, if $\underset{\sim}{\theta}_\omega$ refers to a subsequence for which all θ_i equal a common value θ, then we are back in the situation described at the beginning of the section, in that setting $\overline{R}_\delta^\omega = R_\delta(\theta)$ will suffice. The point, however, is that only subsequences that are of separate interest and that can be expected to occur should be considered, since the reporting of $\overline{R}_\delta^\omega$ is an attempt to model the actual real world performance of δ on a variety of problems. Thus, if C is a confidence procedure, one would like to say that, in repeated actual use, C will fail to contain θ_i no more than $100 \times \overline{R}_C$ % of the time; while if δ is a test, no more than $100 \times \overline{R}_\delta^{\omega_0}$ % of true null hypotheses will be rejected. One can, of course, imagine the "thought experiment" consisting of repeated use of δ for the same θ, but Neyman explicitly created the frequentist theory precisely to eliminate the dependence of statistics on "prior"beliefs or supposed structure about the $\underset{\sim}{\theta}$ that would occur. It is repeatedly stressed in his papers that the measure $\overline{R}_\delta^\omega$ will be valid for <u>any</u> $\underset{\sim}{\theta}_\omega$, and that this is the "breakthrough"provided by frequentist theory.

There are a number of reasons why frequentist theory came to be perceived as simply reporting $(\delta, R_\delta(\theta))$. In the first place, for testing problems where Θ consists of only two points, the null hypothesis (θ_0) and the alternative hypothesis (θ_1), then the ω_0 and ω_1 referred to above could be considered sequences of identical parameters, with $\overline{R}_\delta^{\omega_0} = R_\delta(\theta_0)$ and $\overline{R}_\delta^{\omega_1} = R_\delta(\theta_1)$ corresponding to the actual probabilities of Type I and Type II error. A "natural" generalization to problems with more complicated Θ, would be to consider $R_\delta(\theta)$ in general, forgetting the original motivation. A second reason for the introduction of $R_\delta(\theta)$ was that $\overline{R}_\delta^{\omega}$ is often most easily obtained by finding an upper bound on $R_\delta(\theta)$ (for the type of subsequence $\{\theta_{\omega(i)}\}$ of interest). It is "easiest" to just provide $R_\delta(\theta)$, and let the user of δ infer the appropriate $\overline{R}_\delta^{\omega}$. A third natural reason for consideration of $R_\delta(\theta)$ is for comparison of two procedures δ_1 and δ_2. If $R_{\delta_1}(\theta) < R_{\delta_2}(\theta)$, it will almost invariably follow that δ_1 has better performance than δ_2 in actual long run use. Similarly, for questions of experimental design, $R_\delta(\theta)$ is a basic quantity of interest. It is indeed not surprising that "report $(\delta, R_\delta(\theta))$" came to be perceived as the frequentist viewpoint. Ultimately, however, it is (2.3) (or even better (2.4)) which seems to provide the justification for the frequentist viewpoint, and so it is Definition 1 that we will take as basic.

One final historical matter: it is probably not completely clear who first espoused the frequentist viewpoint as described here (although there is no doubt that it was Neyman who provided the first clear general formalizations). The resolution of this historical matter is complicated by many red herrings, such as significance tests, which are hundreds of years old and can be given a frequentist interpretation. Indeed, a significance test of the null hypothesis that X has distribution P_0, which rejects when a statistic $T(x) > t_0$, can be considered a formal frequentist

procedure with

$$\overline{R}_\delta = P_0(T(X)>t_0).$$

Until Neyman, however, the interpretation of such a procedure seemed to be, upon rejecting, that "either P_0 is false or a very unlikely event has been observed." (See Fisher (1926), for example.) The frequentist interpretation in terms of long run behavior did not become prevalent until after Neyman. (This is not to say that Neyman supported significance testing of a null hypothesis; indeed, he often argued that alternatives must be considered.)

To many, the definition of frequentism being given here may be felt to be too strict. For instance, the quoting of a P-value (in the above setting, $P_0(T(X)>T(x))$) may be felt to be a frequentist procedure by some, since it involves an averaging over the sample space. The reporting of P-values can be given no long run frequency interpretation, in the sense discussed above, however, and cannot even be given such an interpretation in the more general setup of the next section. A P-value actually lies closer to conditional (Bayesian) measures than to frequentist measures (see Berger and Wolpert (1984) for references).

This relates to a point that should be mentioned, namely that the "practicing" frequentist statistician behaves quite differently than the "formal" frequentist defined above, recognizing the rarity of being able to completely state $(\delta, \overline{R}_\delta^\omega)$ (or $(\delta, R_\delta(\theta))$) before experimentation, and thus admitting the need for "adhocery." Presumably, however, such a frequentist attempts to stay as close as possible to the frequentist ideal, so that discussion of the motivation for this ideal is certainly not out of order.

2.2 Conditional Frequentist Approaches

Because of examples such as Example 1, there have long existed conditional

frequentist approaches to statistics. The usual idea is to condition on some event or statistic, such as an ancillary statistic (c.f. Fisher (1956)), and then to do a frequentist calculation.

Example 1 (continued). Defining the ancillary statistic $T(X) = |X_1 - X_2|$, an easy calculation show that

$$P_\theta(C(X) \text{ contains } \theta \mid T(X) = 2) = 1, \text{ and}$$

$$P_\theta(C(X) \text{ contains } \theta \mid T(X) = 0) = \frac{1}{2},$$

corresponding to intuition.

The most comprehensive development of conditional frequentist theory is that in Kiefer (1975, 1976, 1977), Brownie and Kiefer (1977), and Brown (1978). Two, more or less distinct, approaches are discussed in these papers namely "conditional confidence" and "estimated confidence." These two approaches are reviewed in the next two subsections although, since the setting will be that of general L, the approaches will be termed "conditional risk" and "estimated risk."

2.2.1 Conditional Risk

This approach is essentially a formalization of conditioning on an ancillary statistic or "relevant" subset. One considers a partition $\{C^b, b \in B\}$ of $\mathcal{X}$, calculates

$$(2.5) \quad R_\delta^b(\theta) = E_\theta[L(\theta,\delta(X)) \mid C^b],$$

and reports, when $x \in C^b$ is observed, the triple $(\delta(x), C^b, R_\delta^b(\theta))$. The long run frequentist justification for doing this is, of course, that (when C^b has positive probability)

$$(2.6) \quad \lim_{n\to\infty} \frac{\sum_{i=1}^{n} L(\theta, \delta(X_i)) I_{C^b}(X_i)}{\sum_{i=1}^{n} I_{C^b}(X_i)} = R_\delta^b(\theta)$$

with probability one for each θ, so that $R_\delta^b(\theta)$ does measure the average performance

of δ in those problems where X_i falls in C^b (and θ is the same). Furthermore, letting $H(x)$ denote that b for which $x \in C^b$, it is clear that

(2.7) $\quad R_\delta(\theta) = E_\theta R_\delta^{H(X)}(\theta)$,

so that one also has unconditional frequentist validity.

Example 1 (continued). Let $C^1 = \{x: |x_1 - x_2| = 2\}$, and $C^0(x) = \{x: |x_1 - x_2| = 0\}$. Then

$$R_C^1(\theta) = 1 - P_\theta(C(X) \text{ contains } \theta | C^1) = 0, \text{ and}$$

$$R_C^0(\theta) = 1 - P_\theta(C(X) \text{ contains } \theta | C^0) = \frac{1}{2} \ .$$

Example 3a (continued). Let $C^b = \{b, -b\}$ (just the two points) for each $b > 0$. Then

(2.8) $\quad R_\delta^b(\theta) = E_\theta[L(\theta,\delta) | C^b]$

$$= \frac{f(-(\text{sgn}\theta)b|\theta)}{f(-b|\theta) + f(b|\theta)}$$

$$= \frac{1}{1 + \exp\{2b|\theta|\}} \ .$$

Thus, one reports the relevant decision along with $R_\delta^{|x|}(\theta)$, which, when x near zero is observed, will be close to $\frac{1}{2}$ as intuition would suggest. And when x is far from zero, the conditional risk, $R_\delta^{|x|}(\theta)$, will be small. Observe also that

(2.9) $\quad \sup_\theta R_\delta^b(\theta) = 1 \ / \ [1 + \exp(4b)]$.

Although the conditional risk approach has a number of attractive features, as seen in the above examples, it also has several deficiencies. First, there is still present the problem that different θ_i are not present in (2.6). In other words a usable justification, as in Definition 1, cannot be given. This can be corrected if $R_\delta^b(\theta)$ has a usable upper bound $\tilde{R}_\delta^b$, for then, with probability one for all $\underset{\sim}{\theta}$ (under reasonable conditions)

$$(2.10)\quad \limsup_{n\to\infty} \frac{\sum_{i=1}^{n} L(\theta_i, \delta(X_i)) I_{C^b}(X_i)}{\sum_{i=1}^{n} I_{C^b}(X_i)} \leq \tilde{R}_\delta^b \quad ,$$

where the X_i arise from different problems with different θ_i. In Example 1, $R_\delta^b(\theta)$ is constant, so (2.10) automatically obtains. In Example 3a, a usuable bound is given in (2.9), so again one has the validity of (2.10).

The obtaining of a useful bound for $R_\delta^b(\theta)$ is deemed to be of primary importance in Brown (1978), and even in Kiefer (1975) and Brownie and Kiefer (1977) the desirability of choosing $\{C^b\}$ so as to achieve constant $R_\delta^b(\theta)$, or at least a useful upper bound, is stressed. Unfortunately, a useful bound cannot always be achieved. In Example 3b, for instance, it can be shown that, for any partition $\{C^b\}$, $\sup_\theta R_\delta^b(\theta) = \frac{1}{2}$, a useless upper bound.

A second difficulty with conditional confidence is that the choice of the partition $\{C^b\}$ is quite arbitrary, and the justification, even (2.10), depends on this choice. We are not so much referring to the practical difficulty of choosing $\{C^b\}$ (which can be considerable, outside of obvious situations such as Example 1), as to the unappealing arbitrariness in the evaluation of the accuracy of δ that is introduced. In all conditional approaches there will be a certain degree of arbitrariness, but none as extensive as the allowance of arbitrary $\{C^b\}$, especially since the choice of $\{C^b\}$ will rarely have any "outside justification." More about this will be said in Section 2.3.

2.2.2 Estimated Risk

The estimated risk approach replaces the reporting of $(\delta, R_\delta(\theta))$ by the reporting of $(\delta, \hat{R}_\delta(x))$, where $\hat{R}_\delta(x)$ is, in some sense, an estimate of $R_\delta(\theta)$. If, in fact, $\hat{R}_\delta$ is an unbiased estimator of $R_\delta(\theta)$ for all θ, then with probability one (under reasonable conditions),

$$(2.11)\quad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} [L(\theta_i, \delta(X_i)) - \hat{R}_\delta(X_i)] = 0$$

for any sequence $\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots)$. Thus the average performance of δ, in actual repeated use, will be the average of the reported $\hat{R}_\delta(X_i)$. Often an unbiased estimator of $R_\delta(\theta)$ may not exist, or may, in some sense, be undesirable. In such cases, however, one can usually find a desirable function $\hat{R}_\delta(x)$ such that

$$(2.12)\quad E_\theta \hat{R}_\delta(X) \geq R_\delta(\theta) \text{ for all } \theta,$$

in which case (2.11) could be replaced (under reasonable conditions) by

$$(2.13)\quad \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} [L(\theta_i, \delta(X_i)) - \hat{R}_\delta(X_i)] \leq 0,$$

establishing frequentist validity of the report $(\delta, \hat{R}_\delta)$ in the same "optimum" sense as in Definition 1. This is the point to be stressed: the justification for being a frequentist applies equally well to the $(\delta, \hat{R}_\delta)$ report as to the $(\delta, \overline{R}_\delta)$ report (and applies much better than to the $(\delta, R_\delta(\theta))$ report), and yet allows much greater latitude for aligning the report with conditional common sense.

Example 1 (continued). Simply choose

$$(2.14)\quad \hat{R}_\delta(x) = \begin{cases} 0 & \text{if } |x_1 - x_2| = 2 \\ \frac{1}{2} & \text{if } |x_1 - x_2| = 0 \end{cases}$$

and one has an unbiased estimate of risk satisfying the "ideal" (2.11), and providing good conditional reports.

Example 3a (continued). Although an unbiased estimate of risk cannot be obtained, the choice

$$(2.15)\quad \hat{R}_\delta(x) = 1 \,/\, [1 + \exp(4|x|)]$$

satisfies (2.12), and hence (2.13). Thus $(\delta, \hat{R}_\delta)$ provides a (conservatively) valid

frequentist report, and seems much more attractive intuitively than the unconditional error probability bound of .0228.

The above two examples are situations in which $\hat{R}_\delta$ could be chosen to be an appropriate conditional risk or bound on such. The estimated risk approach provides additional flexibility, however, in that it can deal with situations where the conditional risk approach does not provide useful bounds.

Example 3b (continued). It was noted that, for any $\{C^b\}$, $\sup_\theta R_\delta^b(\theta) = \frac{1}{2}$. It is possible, however, to find $\hat{R}_\delta(X)$ such that

$$(2.16) \qquad E_\theta \hat{R}_\delta(X) \geq R_\delta(\theta) = \Phi(-|\theta|) \text{ for all } \theta$$

(where Φ is the standard normal c.d.f.), and such that $\hat{R}_\delta(x)$ decreases to zero as $|x| \to \infty$. Use of such $(\delta, \hat{R}_\delta)$ allows one to report, when $x = 10$ is observed, that the null hypothesis is conclusively rejected. Of come concern here is that any such $\hat{R}_\delta$ must be greater than $\frac{1}{2}$ with positive probability, in order for (2.16) to hold when $\theta = 0$. It would be unappealing to report "I reject H_0 and have estimated error probability of .6." A casual observer would find such a statement most peculiar. Of course, this will tend to arise only when x provides very inconclusive evidence, and hence may not be of much practical concern.

It should be mentioned that the idea of estimated risk definitely precedes Kiefer's work on the subject, although earlier work did not explicitly recognize that use of estimated risk was completely valid from a frequentist perspective. Mention of estimated power can be found in Lehmann (1959), and Sandved (1968) found unbiased estimators of risk for a number of problems. There are also a number of current areas of research where unbiased estimates of risk are commonly

employed, such as Stein estimation.

Example 6. Suppose $X \sim \mathcal{N}_p(\theta, I)$, and that $L(\theta,\delta) = |\theta-\delta|^2$. For $p \geq 3$, Stein (1981) shows that

$$\hat{R}_\delta(x) = p - (p-2)^2 / |x|^2$$

is an unbiased estimator of the risk of

$$\delta(x) = (1 - (p-2) / |x|^2)\, x ,$$

and hence

$$R_\delta(\theta) = E_\theta \hat{R}_\delta(X) < p = R_{\delta^0}(\theta) ,$$

where $\delta^0(x) = x$ is the usual estimator. One can thus report $(\delta(x), \hat{R}_\delta(x))$ with total frequentist justification. (Of course, in this case such a report will be fairly silly for $|x|^2 < (p-2)^2 / p$, since $\hat{R}_\delta$ will then be negative.)

Note that the estimated risk approach can be combined with the conditional risk approach to obtain estimated conditional risk (see Kiefer (1977)). Or one could obtain estimated versions of $\bar{R}_\delta^\omega$ (see Definition 1), valid for subsequences ω of interest. Again, this would increase flexibility, but would result in more ambiguous frequentist justification. Note, on the other hand, that reports of $(\delta, \hat{R}_\delta)$ may require different interpretation than, say, $\bar{R}_\delta^\omega$, as the following example shows.

Example 2b (continued). A possible choice of $\hat{R}_\delta$ would be

(2.17) $\hat{R}_\delta(1) = .5$, $\hat{R}_\delta(2) = .00749$, $\hat{R}_\delta(3) = .00754$.

It can easily be checked that this is an unbiased estimator of risk, and hence the report $(\delta, \hat{R}_\delta)$ is a valid frequentist report. This report also has the attractive frequentist property that, when $x = 1$ is observed, the estimated risk

reported is .5, corresponding to the intuitive (likelihood ratio) assessment. On the other hand, this estimated risk does not have any interpretation as Type I or Type II error, which may be disturbing to some.

2.3 The Choice of a Conditional Frequentist Measure.

The huge variety of possible conditional measures makes choice among them quite difficult. This is especially true for conditional risks $R_\delta^b(\theta)$, and for estimated risks $\hat{R}_\delta^\omega(x)$ where dependence on subsequences ω is allowed. One natural idea is to recognize that a purpose of any version of risk is to communicate some feeling as to the actual loss $L(\theta, \delta)$, and hence introduce a communication loss $L^*(\hat{R}_\delta(x), L(\theta, \delta(x))$ and corresponding communication risk

$$(2.18)\quad R^*(\hat{R}_\delta, \delta, \theta) = E_\theta L^*(\hat{R}_\delta(X), L(\theta, \delta(X))).$$

(Or, of course, one could define these same quantities with $R_\delta^b(\theta)$ or $\hat{R}_\delta^\omega(x)$ in place of $\hat{R}_\delta(x)$.)

Example 4 (continued). In this confidence set situation, a natural choice for L^* is

$$(2.19)\quad L^*(\hat{R}_\delta(x), 1 - I_{C(x)}(\theta)) = (\hat{R}_\delta(x) - [1 - I_{C(x)}(\theta)])^2.$$

(Part of the reason that this would be a sensible measure of how well $\hat{R}_\delta(x)$ communicates whether or not θ is in $C(x)$ is that it is a proper scoring rule. Any proper scoring rule (c.f., Lindley (1982)) would probably serve as well.) Thus, in the situation of Example 1, the unconditional frequentist report $R_\delta(\theta) \equiv \frac{1}{4}$ has

$$R^*(\tfrac{1}{4}, C, \theta) = E_\theta(\tfrac{1}{4} - [1 - I_{C(X)}(\theta)])^2 = \frac{3}{16},$$

while the estimated risk report $\hat{R}_\delta$ given in (2.14) has

$$R^*(\hat{R}_\delta, C, \theta) = E_\theta(\hat{R}_\delta(X) - [1 - I_{C(X)}(\theta)])^2 = \frac{1}{8}.$$

Thus $\hat{R}_\delta$ is uniformly better, as intuition would demand.

An analysis of Example 3(a) similarly shows that $\hat{R}_\delta$, given in (2.15), is better than the best unconditional (Definition 1) statement of $\overline{R}_\delta = .0228$. It is not always true that a sensible $\hat{R}_\delta$ will uniformly dominate an unconditional $\overline{R}_\delta$, however, as Example 3(b) indicates. Here $\overline{R}_\delta = \frac{1}{2}$ can be shown to be better (under squared error L*) for θ near zero than any other $\hat{R}_\delta$ satisfying (2.16), although a decreasing $\hat{R}_\delta(x)$ seems much more satisfying intuitively.

In the above examples, only situations where the conditional or estimated risk was independent of θ (or ω) were considered. Indeed, it is, perhaps, only in such situations that the use of L* makes any sense. To see this, note that there is nothing in the formalism of conditional risk to prevent one from selecting each singleton $\{x\}$ as a conditioning set, so that one gets

$$(2.20) \qquad R^b_\delta(\theta) = E_\theta[L(\theta, \delta(X)) | X = x] = L(\theta, \delta(x)).$$

Clearly this will be optimal from the viewpoint of L*, and is a valid conditional frequentist conclusion. It tends not to be operationally very useful, however.

Attempting to define meaningful criteria, under which to evaluate conditional measures other than the simple $\hat{R}_\delta(x)$, leads to something of a morass. A wide variety of evaluation methods and "admissibility criteria" are proposed in Kiefer (1975, 1976, 1977), Brownie and Kiefer (1977), and Brown (1978). In some sense, the criteria of Brown (1978) are the most appealing, in that they attempt to relate the evaluation of the conditional measure to its likely use, as opposed to trying to determine "intuitively appealing" properties of conditional measures. Our own (somewhat naive or more realistic - take your pick) view of this issue is that one would really like to communicate the posterior expected loss for the

problem, and only evaluation criteria which recognize this will tend to produce reasonable conditional measures. Of course, this probably also applies to the (L*, R*) method for evaluating the simple $\hat{R}_\delta(x)$.

A few final comments are in order concerning the evaluation of $\hat{R}_\delta$. First, it should be emphasized that this is a decision problem with decision space consisting only of $\hat{R}_\delta$ which satisfy the property (2.12). If one removes the restriction (2.12), inadmissibility can result for otherwise admissible $\hat{R}_\delta$. Consider the following example.

Example 7. Suppose $X \sim \mathcal{N}_p(\theta, I)$, where $\theta = (\theta_1,\ldots,\theta_p)$ is unknown, and consider the classical confidence procedure

$$C^0(x) = \{\theta: \ |\theta-x|^2 \leq \chi^2_p(1-\alpha)\} \ ,$$

where $\chi^2_p(1-\alpha)$ is the $1-\alpha$ th percentile of the chi-squared distribution with p degrees of freedom. In the framework of Example 4, this can be considered a frequentist procedure with $\overline{R}_\delta = \alpha$. But with respect to the loss (2.19), Robinson (1979a, 1979b) proves this to be inadmissible for $p = 5$. However, the uniformly better procedure is of the form

$$\hat{R}_{C^0}(x) = \alpha - k(x),$$

where $k(x) > 0$, so that $E_\theta \hat{R}_{C^0}(X) < \alpha$, violating the frequentist validity requirement.

We make no judgement (in this section) as to whether or not a violation of the frequentist validity requirement, in situations such as the above example, is desirable. It is (here) being taken as given that frequentist validity is required, and the exploration concerns the leeway still allowed in the choice of

$\hat{R}_\delta$. For this reason we will also not put L* (or R*) on the same footing as L (or R), even though it is very tempting to do so. In the confidence procedure setting, for instance, one could consider an overall loss such as

$$L(\theta, C(x), \hat{R}_C(x)) = k_1[1 - I_{C(x)}(\theta)] + k_2[\hat{R}_C(x) - (1 - I_{C(x)}(\theta))]^2 ;$$

which recognizes not only the importance of having C(x) contain θ but also the importance of properly communicating whether or not it does. Decision theory with such loss functions would be quite interesting, but it is probably best to keep L and L* separate. Also, it is probably unwise to attempt to follow the road much farther, i.e., to actually report (or estimate) R* as if it were a meaningful quantity (as opposed to a device for selection of $\hat{R}_\delta$). This is mainly a feeling based on the seeming inevitability of Bayesian analysis as a mechanism for sensible communication of whether or not θ is in C(x) (c.f., Lindley (1982)). To a non-Bayesian, however, there may be some appeal to reporting (δ, $\hat{R}_\delta$, R*), and perhaps there is merit in so doing. This is especially plausible when the "communication" aspect can be considered as part of the real problem.

One final comment about the situation of Example 7 is in order. A recently much studied problem has been the replacement of the usual confidence spheres $\{C^0(x)\}$ by spheres $\{C^*(x)\}$ of the same size, but centered at the Stein-type estimator

$$\delta^{J\text{-}S}(x) = (1 - c / |x|^2)^+ x.$$

For appropriate c (depending on p) it can be shown (c.f., Hwang and Casella (1982)) that

$$R_{C^*}(\theta) = 1 - P_\theta(C^*(X) \text{ contains } \theta) < \alpha$$

for all θ. But $\overline{R}_{C^*} = \sup_{\theta} R_{C^*}(\theta) = \alpha$, so that one can not improve on C^0 in terms of an unconditional frequentist conclusion, independent of θ. It is clearly possible, however (though probably difficult), to find $\hat{R}_{C^*} < \alpha$ such that

$$R_{C^*}(\theta) \leq E_\theta \hat{R}_{C^*}(X) < \alpha,$$

allowing an estimated risk report of $(C^*, \hat{R}_{C^*})$ which offers a clear gain.

3. Conflicts Between Frequentist and Conditional Analysis

3.1 Introduction

The examples in Section 1 made it clear that unconditional frequentist theory could not be suitable as a general philosophy of statistics. In Section 2 it was seen, however, that conditional frequentist theory was just as valid from the frequentist viewpoint, and seemed to offer much greater scope for correspondence with "conditional common sense." The obvious issue, therefore, is whether or not the conditional frequentist theory is itself rich enough to provide a suitable general philosophy. There are, disturbingly, still simple examples indicating concern as to this suitability. For instance, consider the following modification of Example 2b.

Example 2c. Suppose the probability structure is

	X = 1	2	3
P_0	.05	.15	.8
P_1	.051	.849	.1

Now the risks (error probabilities) of the test δ, which rejects when $x = 1$ or 2, are $R_\delta(0) = .2$ and $R_\delta(1) = .1$. Conditional risk theory will never work well for

such a situation, since one of the sets C^b, of the partition, must contain only one point, resulting in a conditional risk of zero or one (unless an extremely artificial randomization is introduced). And, since $R_\delta(0)$ and $R_\delta(1)$ differ substantially, the estimated risk theory does not easily apply. Perhaps some version of the estimated risk theory which allows dependence on ω (or θ) would give sensible answers (by which is meant, not only frequentist validity, but also an appropriate expression of doubt for $x = 1$), but it is not easy to find such.

Although the difficulty in successfully dealing with simple examples, such as Example 2c, via frequentist theory, is a cause for concern, the sheer vastness of the conditional frequentist domain makes unlikely a "disproof by counter-example." Serious axiomatic conflicts with frequentist or conditional frequentist viewpoints exist, however, and it is to a brief discussion of these that we now turn. Section 3.2 briefly reviews the axiomatics leading to the likelihood principle, and the resultant conflict with frequentist ideas. Section 3.3 discusses one of the most important of these conflicts, that concerning the role of the stopping rule in statistical analysis. Indeed, because of the importance and intuitive difficulties concerning this latter issue, a separate argument is given, showing the incompatibility of frequentist admissibility with the idea that the stopping rule must be taken into account.

3.2 The Likelihood Principle

We will forgo extensive discussion of the Likelihood Principle (LP) here, presenting only a bare bones outline of its implications and its axiomatic development (due to Birnbaum (1962)). A recent monograph (Berger and Wolpert (1984)) extensively discusses the LP, its history, and its ramifications.

In what follows, $\mathcal{X}$ will be considered to be discrete, so that the likelihood function $f(x|\theta)$, i.e., the density considered as a function of θ for the actual x, is well defined. This restriction can be removed; indeed, in Berger and Wolpert (1984) a generalization of the LP (called the relative likelihood principle) is developed which yields the same consequences as the LP and yet does not require the existence of densities (or even parametric models). Thus the "common criticism" that the LP is not valid, because it does not apply to situations where the model is uncertain, is not applicable to the appropriate generalization of the LP. (And it can even be argued that, since in reality all $\mathcal{X}$ are discrete and finite and for such $\mathcal{X}$ all families of distributions are parametric - the most general possible index θ being simply the vector of probabilities of the elements of $\mathcal{X}$, the simple version of the LP always applies. Such an argument is given in Basu (1975).) This is not the place to extensively discuss all the criticisms of the LP that have been raised. (See Berger and Wolpert (1984) and Berger (1984a) for such discussion.) We merely wish to make the point that Birnbaum's axiomatic development should be taken seriously, and can not be easily dismissed.

Following Birnbaum's notation, we let E be an experiment consisting of observing $X \sim P_\theta$, and are concerned with the "evidence" or "information" about θ that is obtained (or should be reported) upon observing x. This will be denoted Ev(E,x). (This "evidence" could be anything at all, including one or several frequentist measures: note that by listing E we allow "Ev" to depend on full knowledge of all aspects of the experiment, and not just the observed x.)

The Likelihood Principle. Ev (E,x) should depend on E and x only through the likelihood function, $f(x|\theta)$, for the observed x. Two likelihood functions for

(the same unknown) θ yield identical evidence about θ if they are proportional (as functions of θ).

Example 2d. Again assume $\mathcal{X} = \{1,2,3\}$ and $\Theta = \{0,1\}$, and consider experiments E_1 and E_2 which consist of observing X_1 and X_2 with the above $\mathcal{X}$ and Θ, but with probability densities as follows:

X_1

	1	2	3
$f_1(x_1\|0)$	.9	.05	.05
$f_1(x_1\|1)$	.09	.055	.855

,

X_2

	1	2	3
$f_2(x_2\|0)$	.26	.73	.01
$f_2(x_2\|1)$	.026	.803	.171

.

If, now, $x_1 = 1$ is observed, the LP states that $Ev(E_1,1)$ should depend on the experiment only through $(f_1(1|0), f_1(1|1)) = (.9,.09)$. Furthermore, since this is proportional to $(.26,.026) = (f_2(1|0), f_2(1|1))$, it should be true that $Ev(E_2,1) = Ev(E_1,1)$. (Another way of stating the LP for testing hypotheses, as here, is that $Ev(E,x)$ should depend on E and x only through the likelihood ratio for the observed x.) It is similarly clear that, according to the LP, $Ev(E_1,2) = Ev(E_2,2)$ and $Ev(E_1,3) = Ev(E_2,3)$. Hence, no matter which experiment is performed, the same evidentiary conclusion about θ should be reached for the given observation. (This example is given in Berger and Wolpert (1984).)

The above example clearly indicates the startling nature of the LP. Experiments E_1 and E_2 are very different from a frequentist perspective. For instance, the decision procedure which decides $\theta = 0$ when the observation is 1 and decides $\theta = 1$ otherwise is a most powerful test with error probabilities (of Type I and Type II, respectively) .10 and .09 for E_1, and .74 and .026 for E_2. Thus the classical frequentist would report drastically different "evidence"

from the two experiments. (And conditional frequentist approaches are very unlikely to give similar conclusions: indeed, for E_2 it is very hard to perform any sensible conditional frequentist analysis, because of the three point $\mathcal{X}$ and the widely differing error probabilities.)

This example emphasizes a very important issue. It is clear that experiment E_1 is more likely to provide useful information about θ, as reflected by the overall better error probabilities. The LP, in no sense, contradicts this. Indeed, the LP says nothing about experimental design or any other situation involving an evaluation for not yet observed X. The LP applies only to the information <u>about θ</u> that is available from knowledge of E and the <u>observed x.</u> Even though E_1 has a much better chance of yielding good information, the LP states that the conclusion, once x is at hand, should be the same, regardless of whether x came from E_1 or E_2. The conflict of the LP with frequentist justifications seems inescapable. (See also Birnbaum (1977).)

A committed frequentist might look at this example and reject the LP out of hand, although some unease will undoubtedly be present because of the equal likelihood ratios in the experiments. Very troubling, however, is the fact that the LP is a direct consequence of two other "obvious" principles, the Sufficiency Principle and the Weak Conditionality Principle.

<u>The Sufficiency Principle.</u> If T is a sufficient statistic for θ in an experiment E, and $T(x_1) = T(x_2)$, then $Ev(E,x_1) = Ev(E,x_2)$.

<u>The Weak Conditionality Principle.</u> Let E_1 consist of observing X_1 with density $f_1(x_1|\theta)$ and E_2 consist of observing X_2 with density $f_2(x_2|\theta)$. (Here θ is the <u>same</u> quantity in each experiment.) Consider the <u>mixed</u> experiment E consisting of observing J = 1 or 2 with probability $\frac{1}{2}$ each (independent of everything - say, the result of a fair coin flip), and then performing experiment E_J. The random

quantity observed from E is thus (J,X_J). Then it should be true that

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

i.e., the evidence obtained about θ is simply the evidence from the experiment actually performed.

Almost everyone accepts the Sufficiency Principle, and the Weak Conditionality Principle (essentially due to Cox (1958)) seems very natural, being the weakest form of conditioning imaginable. (Some frequentists might reject the Weak Conditionality Principle by essentially rejecting the idea that the goal is to communicate evidence about θ. If the goal is simply to determine repeated performance of a procedure in use, then the repeated performance for E will likely differ from the repeated performance for one of the E_j. It seems unlikely that such a view could ever gain wide acceptance, however; insisting on reporting 75% coverage in Example 1, for instance, is hardly tenable.) Birnbaum's surprising result was that this weakest form of conditioning (together with sufficiency) implies that complete conditioning, down to $f(x|\theta)$, should be done.

Theorem (Birnbaum (1962)). The Sufficiency Principle and the Weak Conditionality Principle together imply the LP.

All the generalizations of the LP that were referred to earlier also follow from sufficiency and weak conditionality, and so a frequentist is left with the uncomfortable choice of rejecting sufficiency or weak conditionality. It is a conflict which clearly deserves careful thought.

This is not to say that all frequentist procedures violate the LP. In fact, it is well known that a large portion of standard frequentist procedures can be interpreted as Bayesian procedures with "noninformative" priors and hence are consistent with the LP. (Bayesian procedures always depend on E and x only through $f(x|\theta)$.) The frequentist concept of evidence, based on some type of average over X, is clearly a concept in conflict with the LP, however, and Example 2d shows how dramatic the conflict can be.

3.3 The Stopping Rule Principle

One of the most important practical applications of the LP is the Stopping Rule Principle (SRP), developed (at various levels) in Barnard (1947), Birnbaum (1962), Pratt (1965), and Berger and Wolpert (1984). Suppose E is a sequential experiment, with possible observations $X_1, X_2, \ldots$ having probability distribution P_θ (determined by the finite dimensional distributions, of course), with θ unknown. For convenience, only non-randomized stopping rules, τ, are considered. Such a stopping rule can, most conveniently, be represented by a sequence of sets $\{(A_n, B_n)\}$, where

(3.1) if $\underset{\sim}{x}^n = (x_1, \ldots, x_n) \in A_n$, stop sampling;
if $\underset{\sim}{x}^n \in B_n$, continue sampling.

(Without loss of generality, it can be assumed that, if $\underset{\sim}{x}^n \in A_n$, then $\underset{\sim}{x}^j \in B_j$ for all $j < n$.) Let N denote the stopping time (i.e., the n for which $\underset{\sim}{x}^n \in A_n$). Only proper stopping rules (i.e., those for which $P_\theta(N<\infty) = 1$) will be considered.

Stopping Rule Principle. For a sequential experiment E with observed data $\underset{\sim}{x}^n$, $Ev(E, \underset{\sim}{x}^n)$ should not depend on the stopping rule τ.

In words, the SRP simply states that the reason for stopping sampling should

be irrelevant to evidentiary conclusions about θ (providing this reason, as above, does not depend on θ in any fashion, except indirectly through the x_i). One can constantly monitor incoming data, and stop at any time that the data looks good enough (or bad enough or whatever), and this "optional stopping" should play no role for a valid measure of evidence.

The practical implications of the SRP are enormous, since optional stopping is a huge problem in many areas of practical statistics, such as clinical trials. Scrupulous experimenters and scientists would delight in the freedom to stop an experiment at any point felt to be appropriate, without having to worry about an effect of such optional stopping. And the scientific community, as a whole, would no longer be prey to misleading (frequentist)conclusions arising from situations where optional stopping was employed, but not reported.

Of course, the hitch in all this is that frequentist measures are very dependent on the stopping rule, and can not be used in conjunction with the SRP. Indeed, frequentist intuition will generally react to the SRP with outrage, it being "obvious" that "stopping when the data looks good" will bias the results (and, of course, it will in a frequentist sense). Since, however, the SRP can be shown to be a trivial consequence of the LP (or of the relative likelihood principle of Berger and Wolpert (1984), if complete generality is desired), a major conflict in intuition again surfaces; rejecting the SRP corresponds to rejecting either sufficiency or weak conditionality.

In an attempt to resolve the issue (in favor of the SRP) it is interesting to observe that the frequentist intuition that "the stopping rule matters" is itself inconsistent with the frequentist concept of admissibility. Consider the following example.

Example 8. Suppose $\Theta = R^1$, and that it is desired to estimate θ under squared error loss, $L(\theta,\delta) = (\theta-\delta)^2$. Imagine, now, two possible stopping rules, τ_1 and τ_2, determined by $\{(A_n^1, B_n^1)\}$ and $\{(A_n^2, B_n^2)\}$, respectively, and suppose that, for some n_0, there exists a set $A \subset A_{n_0}^1 \cap A_{n_0}^2$ such that $P_\theta(A) > 0$ and on which the estimators δ_1 and δ_2, that would be used under τ_1 and τ_2, respectively, are different. (If no such A exists, then the stopping rules are not really having any effect on the decision.)

To see that this conflicts with admissibility, or more basically with long-run frequentist optimality, imagine that one will be faced with a series of such experiments, in half of which τ_1 will be used, and in half of which τ_2 will be used. Then it is a simple matter to show that one could do better (for a sequence of θ_i such that $P_{\theta_i}(A) > \varepsilon > 0$) by using the estimator δ_j if τ_j is used and $N \neq n_0$ or $\underset{\sim}{x}^{n_0} \notin A$, while using

$$(3.2) \qquad \frac{1}{2}\delta_1(\underset{\sim}{x}^{n_0}) + \frac{1}{2}\delta_2(\underset{\sim}{x}^{n_0}) \text{ if } N = n_0 \text{ and } \underset{\sim}{x}^{n_0} \in A.$$

A formal statement of this would involve, for instance, the consideration of the mixed experiment E, consisting of observing J = 1 or 2 with probability $\frac{1}{2}$ each and then doing the sequential experiment with stopping rule τ_J. This E is a well defined sequential experiment with observation $(J,\underset{\sim}{X}^N)$ (N being the implied stopping time for E), and it is trivial to show that a sufficient statistic for θ (in the experiment E) is

$$T((j,\underset{\sim}{x}^n)) = \begin{cases} \underset{\sim}{x}^n \text{ if } n = n_0,\ \underset{\sim}{x}^n \in A \\ (j,\underset{\sim}{x}^n) \quad \text{otherwise.} \end{cases}$$

If, now, "the stopping rule does matter" then one would presumably use δ_J to estimate θ, but, by construction, this is not a function of the sufficient statistic alone. Hence, since the loss is strictly convex, Rao-Blackwellization of the estimator (via (3.2)) would result in an estimator with strictly better frequentist risk. Thus, admissibility (or long run frequentist validity) implies that the stopping rule should be ignored in making the decision (at least for the $\underset{\sim}{x}^n$ that can be observed under either stopping rule).

4. Conclusions

Some general conclusions seem possible from the preceeding discussions. The first is that the issue of conditioning is serious, and deserves careful consideration by all statisticians. It is a tribute to Kiefer's unswerving pursuit of scientific truth that he recognized this issue and sought to resolve it, even though the issue is an uncomfortable one for frequentists.

The second conclusion that (at least tenuously) can be reached is that, even though conditional frequentist approaches can go a very long way towards achieving compatibility with the conditional view, complete reconciliation appears to be impossible (see also Birnbaum (1977)). One thus has an uncomfortable choice: either abandon frequentist justification as an absolute must, or resign oneself to the possibility that, from time to time, one will be forced to state a conclusion that is at variance with conditional common sense.

Neyman and Kiefer made the second choice. I have made the first choice, essentially because of a refusal to be put in the position of having to give a conclusion for a real statistical problem which I know is (conditionally) a silly conclusion. This is not to say that the frequentist view can not be of great usefulness, but does say that, as a philosophical foundation for statistics, I find it unsuitable.

This raises the issue of what should serve as a philosophical foundation of statistics, and the "obvious" answer is Bayesian analysis, which seems to be the only approach capable of guaranteeing sensible conditional answers (c.f., Berger (1983, 1984a, 1984b) and Berger and Wolpert (1984)). The practical issue of (often extreme) uncertainty in prior knowledge then raises its head, however, along with such issues as the need for "scientific objectivity" and the practical difficulties in complicated situations of obtaining any answer at all. The value of frequentist calculations can be considerable for these and other reasons, as

discussed in the above mentioned articles (which also have earlier references). Indeed it could be argued that, as a practical matter, the frequentist viewpoint will tend to give better answers than the Bayesian viewpoint, even if the latter is philosophically correct. Having now spent a number of years attacking problems from both perspectives, I feel nearly certain that this is not the case, but a discussion of these matters would clearly take us too far afield.

It should also be observed that the entire discussion in this paper has been directed towards the statistical conclusion that will be made once the data is at hand. The problem of designing good experiments, dependent on knowing the expected performance of procedures that will be used, clearly involves a strong frequentist component (although it can be argued that more attention should be paid to Bayesian matters, here, also). Indeed, the frequentist viewpoint partly arose as an effort to unify these two aspects of statistics (see Pearson (1962)). There is no questioning the immense contributions to statistics that have been made by Neyman and Kiefer by adopting this frequentist viewpoint. It seems reasonably clear, however, that a grand unification of the design and evidentiary aspects of statistics, under the frequentist banner, is impossible.

Acknowledgments

Of great help in my struggle to understand these issues were Larry Brown, Leon Gleser, Bruce Hill, Ker-Chau Li, Herman Rubin, and, of course, Jack Kiefer.

References

Barnard, G. A. (1947). A review of "Sequential Analysis" by Abraham Wald. J. Amer. Statist. Assoc. 42, 658-669.

Basu, D. (1975). Statistical information and likelihood (with Discussion). Sankhyā, Ser. A 37, 1-71.

Berger, J. (1983). The robust Bayesian viewpoint. In Robustness in Bayesian Statistics (J. Kadane, ed.). North-Holland, Amsterdam.

Berger, J. (1984a). In defense of the likelihood principle: axiomatics and coherency. In: Bayesian Statistics II (J. M. Bernardo, M. H. DeGroot, D. Lindley, and A. Smith, eds.).

Berger, J. (1984b). Bayesian salesmanship. In: Bayesian Inference and Decision Techniques with Applications: Essays in Honor of Bruno deFinetti (P. K. Goel and A. Zellner, eds.). North-Holland, Amsterdam.

Berger, J. and Wolpert, R. (1984). The Likelihood Principle: A Review, Generalizations, and Statistical Implications. To appear in the Monograph Series of the Institute of Mathematical Statistics.

Birnbaum, A. (1962). On the foundations of statistical inference (with Discussion). J. Amer. Statist. Assoc. 57, 269-306.

Birnbaum, A. (1977). The Neyman-Pearson theory as decision theory and as inference theory: with a criticism of the Lindley-Savage argument for Bayesian theory. Synthese 36, 19-49.

Brown, L. D. (1978). A contribution to Kiefer's theory of conditional confidence procedures. Ann. Statist. 6, 59-71.

Brownie, C. and Kiefer, J. (1977). The ideas of conditional confidence in the simplest setting. Commun. Statist. - Theor. Meth. A 6 (8), 691-751.

Cox, D. R. (1958). Some problems connected with statistical inference. Ann. Math. Statist. 29, 357-372.

Fisher, R. A. (1926). The arrangement of field experiments. J. Ministry Agric. Great Brit. 33, 503-513.

Fisher, R. A. (1956). Statistical Methods and Scientific Inference. Oliver and Boyd, Edinburgh.

Hwang, J. T. and Casella, G. (1982). Minimax confidence sets for the mean of a multivariate normal distribution. Ann. Statist. 10, 868-881.

Kiefer, J. (1975). Conditional confidence approach in multi-decision problems. In: Multivariate Analysis IV (P. R. Krishnaiah, ed.). Academic Press, New York.

Kiefer, J. (1976). Admissibility of conditional confidence procedures. Ann. Math. Statist. 4, 836-865.

Kiefer, J. (1977). Conditional confidence statements and confidence estimators (with Discussion). J. Amer. Statist. Assoc. 72, 789-827.

Lehmann, E. L. (1959). Testing Statistical Hypotheses. Wiley, New York.

Lindley, D. V. (1982). Scoring rules and the inevitability of probability. Int. Statist. Rev. 50, 1-26.

Neyman, J. (1957). 'Inductive behavior' as a basic concept of philosophy of science. Rev. Intl. Statist. Inst. 25, 7-22.

Neyman, J. (1967). A Selection of Early Statistical Papers of J. Neyman. University of California Press, Berkeley.

Pearson, E. S. (1962). Some thoughts on statistical inference. Ann. Math. Statist. 33, 394-403.

Pratt, J. W. (1965). Bayesian interpretation of standard inference statements (with Discussion). J. Roy. Statist. Soc. B 27, 169-203.

Robinson, G. K. (1979a). Conditional properties of statistical procedures. Ann. Statist. 7, 742-755.

Robinson, G. K. (1979b). Conditional properties of statistical procedures for location and scale parameters. Ann. Statist. 7, 756-771.

Sandved, E. (1968). Ancillary statistics and estimation of the loss in estimation problems. Ann. Math. Statist. 39, 1755-1758.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. Ann. Statist. 9, 1135-1151.