CHAPTER 4. CONSEQUENCES AND CRITICISMS OF THE LIKELIHOOD PRINCIPLE AND RELATIVE LIKELIHOOD PRINCIPLE

Most people who reject the LP do so because it has consequences they do not like. Of course any theory deserves to be rejected if its consequences are erroneous, but great care must be taken in making sure that the consequences really are wrong and not just in opposition to the intuition currently dominant in the field. In this section we discuss some of the more surprising consequences of the LP and RLP, and investigate the conflicts with prevalent statistical intuition. It will come as no surprise that we feel that the conflicts are always resolved in favor of the LP and RLP.

4.1 INCOMPATIBILITY WITH FREQUENTIST CONCEPTS

4.1.1 Introduction

The philosophical incompatibility of the LP and the frequentist viewpoint is clear, since the LP deals only with the observed x, while frequentist analyses involve averages over possible observations. It cannot be said, however, that any particular frequentist procedure conflicts with the LP, since the procedure could happen to correspond to a sensible conditional procedure. Such a correspondence does, in fact, occur in many statistical situations. For instance, much of frequentist normal distribution theory inference provides the same numerical measures of "confidence" as does noninformative prior conditional Bayesian theory (because of the symmetries or group structure of the problem), although the interpretations of these measures are different. (A cynic might argue that frequentist statistics has survived precisely because of such lucky correspondences.) Nevertheless, enough direct conflicts have been

(and will be) seen to justify viewing the LP as revolutionary from a frequentist perspective.

We have already alluded to the fact that a frequentist can logically dismiss the LP, essentially by rejecting the WCP and concluding that the concept of learning or drawing conclusions about θ, for a particular experiment, is meaningless. Thus Neyman (c.f. Neyman (1957, 1967, 1977)) espouses the viewpoint that only the performance of a procedure in repeated use is relevant, and that it is a mistake to think in terms of learning about particular θ. Though logically viable, this viewpoint is scientifically unappealing. Experiments are done precisely to obtain "evidence" about unknown θ, and investigators will not take kindly to being told that this is meaningless. Thus Birnbaum (1977) argues that Neyman-Pearson conclusions are virtually always used in an "evidentiary" fashion, rather than as measures of procedure performance in repeated use. Savage put this very succinctly when talking about confidence sets in Savage et. al. (1962):

> "The only use I know for a confidence interval is to have confidence in it."

Supposing then that we are going to use a frequency measure as a measure of evidence about θ, what classical justifications for such behavior can be advanced? There are at least the following four:

(i) Frequency measures are "objective", having a well defined physical interpretation, and science demands objective statistical measures.

(ii) The use of frequency measures (and procedures based on them) is reasonably sound and safe for nonspecialists.

(iii) One needs "repeatable" experiments in science, i.e., any evidence gathered about θ should also be likely to be found if the experiment is repeated; this will supposedly be true if frequency measures of evidence are used.

(iv) The following principle should be followed:

CONFIDENCE PRINCIPLE. Any statistician who uses a methodology in which he makes

statements or draws conclusions with specified accuracy, should be guaranteed that in the long run his actual accuracy will be at least that promised.

We will briefly examine these four justifications.

4.1.2 Objectivity

It should be observed, first of all, that the LP is entirely objective, stating only that the evidence about θ is contained in the likelihood function. Also, the likelihood function has as much physical reality as any frequency measure calculated for a presumed model. It would thus be logically sound to pass on to the next issue. We dally, however, because of the problem of using the likelihood function. Indeed, since in Chapter 5 we will argue for Bayesian use of the likelihood function, issues of objectivity will become relevant.

The Bayesian answers to criticisms of objectivity are either (i) objectivity is a myth, or (ii) only through "noninformative" prior Bayesian analysis can objectivity be really attained. As an example of the first argument, Box (1980) states:

> "In the past, the need for probabilities expressing prior belief has often been thought of, not as a necessity for all scientific inference, but rather as a feature peculiar to Bayesian inference. This seems to come from the curious idea that an outright assumption does not count as a prior belief... I believe that it is impossible logically to distinguish between model assumptions and the prior distribution of the parameters."

A general review of this objectivity issue is given in Berger and Berry (1987). (See also Berger (1985).) The only portion of frequentist theory formally exempt from the argument is (completely) nonparametric analysis, and, even then, the choice of a particular procedure to use can be argued to be a highly subjective input.

If the model can be claimed to have some objective status, there is still argument (ii) (above) to contend with. The idea behind this argument is that one can lay claim to objectivity only by purposely striving for it, through use of what is deemed to be an "objective prior." Substantial support for this position can be found in Jeffreys (1961), Box and Tiao (1973), Zellner (1971), Rosenkranz (1977), Bernardo (1979), Berger (1980,1984e), and Jaynes (1981, 1982). Regardless of the validity of argument (ii), it is a fact that use of noninformative priors is objective, purposely not involving subjective prior opinions, and is consistent with the LP. The measures of evidence used are, of course, probabilistic statements about the unknown θ itself (through the formal posterior distribution of θ) and hence may be deemed less "real", but a very strong case can be made that "evidence" about uncertain quantities should only be quantified probabilistically (c.f. deFinetti (1972, 1974)). There are also other likelihood based methods which can be classified as objective, as will be seen in Chapter 5. Hence, even if deemed obtainable and desirable, objectivity is not a reason to reject the LP in favor of frequency measures.

4.1.3 Procedures for Nonspecialists

We accept the argument that it is important to develop reasonably simple statistical procedures which can be safely used by nonspecialists. However, it is not at all clear that this need be done from a frequency viewpoint. First, frequency methods often attain formal simplicity by obscuring difficult issues, such as the choice of error probabilities in a test or the choice of a partition in a conditional confidence procedure (see Section 2.5). Second, relatively simple procedures and methods of evaluation consistent with the LP can be developed (w/o the introduction of subjective priors) as the books of Jeffreys (1961), Box and Tiao (1973), and Zellner (1971) indicate. We are continually surprised at the ease with which the use of noninformative priors, as in these books, gives excellent (conditional) procedures. Indeed, as mentioned earlier, many reasonable

frequentist procedures are, at least approximately, noninformative prior Bayes procedures, and "frequency confidence" then often coincides with "posterior confidence." When this correspondence does not occur, such as in unconditional frequentist approaches to the examples in Section 2.1, the frequentist approach is definitely suspect. Further discussion and references can be found in Berger (1980). Note that we are not maintaining that the use of noninformative priors solves all problems and is foolproof, but only that, if procedures which are simple to use and interpret are deemed necessary, then there are good conditional alternatives to frequentist development of procedures. We have also slighted the subjective Bayes solution to the problem, which will, however, be discussed in Chapter 5.

In this situation, where a procedure is developed for use by nonspecialists, the performance of the procedure in repeated use is certainly relevant (see Section 3.5.4), though not necessarily of primary importance. Good frequency performance can even be of interest to the strict conditionalist, as the following example indicates.

EXAMPLE 16. Suppose a confidence procedure $C(x)$ is to be used (i.e., when $X = x$ is observed it will be stated that $\theta \in C(x)$), having frequentist coverage probability

$$\Gamma(\theta) \equiv P_\theta(C(X) \text{ contains } \theta) \geq 1-\alpha.$$

A conditional Bayesian (for simplicity) would, for a prior distribution π on Θ, be interested in having good posterior probability that θ is in $C(x)$, i.e., would want

$$\lambda(x) \equiv P^{\pi(\theta|x)}(\theta \in C(x))$$

to be large, where $\pi(\theta|x)$ is the posterior probability distribution of θ given x. But, letting m denote the marginal distribution of X (i.e., $m(\cdot) = E^\pi P_\theta(\cdot)$) and $I_B(y)$ denote the usual indicator function on a set B, it is clear that

$$
\begin{aligned}
E^m\lambda(X) &= E^m P^{\pi(\theta|X)}(\theta \in C(X)) \\
&= E^{\text{joint distbn. } (\theta,X)}[I_{C(X)}(\theta)] \\
&= E^{\pi} P_{\theta}(C(X) \text{ contains } \theta) \\
&\geq 1-\alpha.
\end{aligned}
$$

Since this relationship holds regardless of π, a conditionalist could feel that $\lambda(x)$ is "likely" to be large if $C(x)$ is used and α is small, and hence be willing to use $C(x)$ when unable to carry out a trustworthy Bayesian analysis. See Pratt (1965) and Berger (1984b) for more general development and specific examples.

It is important to emphasize that the primary goal in situations such as Example 16 should still be good conditional performance, and that the frequentist measure does not guarantee this. Conceivably, $\lambda(x)$ could be very small for some x (and all m), which is certainly relevant since such x could be observed. Thus our view is that procedures should usually be developed from a conditional viewpoint, and their frequency properties perhaps investigated to ensure robustness. Of course the already existing classical procedures which have good conditional properties are fine. Other discussions of this point can be found in Godambe and Thompson (1977), Godambe (1982a,b), and Berger (1984e).

4.1.4 Repeatability

There is certainly truth to the observation that, if a scientific experiment claims to have obtained strong evidence about θ, then many scientists expect future similar experiments to also provide strong evidence. The frequency measures, based on imagining repetitions of the experiment, seem ideally suited to achieve this. There is a serious concern here, however, as the following example indicates.

EXAMPLE 17. Suppose X has the two point distribution given by $P_\theta(X = 0) = .99$ and $P_\theta(X = \theta) = .01$. (Either θ will be measured exactly, or no observation will be recorded.) If now $x = 5$ is observed, it should certainly be concluded

that $\theta = 5$ exactly (very strong "evidence"), but repetitions of the experiment are very unlikely to reproduce the result.

It could perhaps be argued that science should not believe "lucky" observations, as in the previous example, and hence should not think conditionally on the data. This seems too severe a straightjacket, however. One can always be skeptical of lucky observations and seek possible alternative reasons for them, but their conditional evidential interpretation should be allowed. Such conditional interpretations can, of course, also be verified or disproved by future investigations.

4.1.5 The Confidence Principle

The Confidence Principle was implicit in much of Neyman's early development of the frequentist viewpoint (c.f. Neyman (1967) and also Neyman (1957, 1977) and Berger (1984c)), and was stated explicitly by Birnbaum (c.f. Giere (1977) and Birnbaum (1968, 1970, 1977)), who ultimately came to reject the LP because of its conflict with the Confidence Principle. Other discussions of this or related principles can be found in Cox and Hinkley (1974) (which distinguishes between strong and weak versions, the weak version allowing conditioning on relevant subsets), Kiefer (1977b), Le Cam (1977), and Barnard and Godambe (1982). Critical discussion can be found in Jeffreys (1961), Hacking (1965), Edwards (1972), deFinetti (1972, 1974), Pratt (1977), and Jaynes (1981, 1982). The following mathematical formulation of the Confidence Principle will be useful in the discussion, and is related to the Evaluation Game in Section 3.7.2.

THE FORMAL CONFIDENCE PRINCIPLE. A procedure δ is to be used for a sequence of problems consisting of observing $X_i \sim P_{\theta_i}$. A criterion, $L(\theta_i, \delta(x_i))$, measures the performance of δ in each problem (small L being good). One should report, as the "confidence" in use of δ,

$$\bar{R}(\delta) = \sup_{\underset{\sim}{\theta}} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} L(\theta_i, \delta(x_i)), \tag{4.1.1}$$

assuming the limit exists with probability one. (It can usually be shown that $\bar{R}(\delta) = \sup_{\theta} R(\theta,\delta)$, *where* $R(\theta,\delta) = E_{\theta}L(\theta,\delta(X))$.)

EXAMPLE 18. Suppose δ is a confidence procedure, so that $\delta(x_i) \subset \Theta$ will be the confidence set when x_i is observed. The natural measure of the performance of $\delta(x_i)$ is

$$L(\theta_i,\delta(x_i)) = 1\text{-}I_{\delta(x_i)}(\theta_i),$$

since this measures whether or not $\delta(x_i)$ *does* contain θ_i. The risk of δ is

$$R(\theta,\delta) = E_{\theta}L(\theta,\delta(X)) = 1\text{-}P_{\theta}(\delta(X) \text{ contains } \theta),$$

and it is easy to show, for this problem, that

$$\bar{R}(\delta) = \sup_{\theta} R(\theta,\delta) = 1\text{-}\inf_{\theta} P_{\theta}(\delta(X) \text{ contains } \theta).$$

Hence the "report," according to the Confidence Principle, should be one minus the minimum coverage probability of δ.

Although the Confidence Principle is formulated above only in terms of repetitive use of δ for problems of the same form (but possibly differing θ_i), it can easily be generalized to include use of δ for different types of problems. Such a generalization adds little conceptually, however. The appeal of the Confidence Principle is undeniable. By following it, the *actual* average performance of δ in repeated use will be at least as good as the *reported* performance $\bar{R}(\delta)$. There are several problems in following the Confidence Principle, however.

The first difficulty is that, in virtually all statistical investigations, extensive assumptions concerning the model, etc., are made. Thus a person claiming to err no more than 5% of the time because he follows the Confidence Principle, is really saying he errs no more than 5% of the time if all the model assumptions he makes are correct. This removes some of the lustre from the principle.

A second serious issue is the need to have a valid bound, $\bar{R}(\delta)$, on the performance of δ. This is an often unappreciated aspect of the frequentist position. Indeed, the frequentist position is often viewed as requiring

only the reporting of the function $R(\theta,\delta)$. Without the bound, $\bar{R}(\delta)$, however, no guarantee of long run performance, in actual use of δ on different problems, can be given.

EXAMPLE 19. Consider simple versus simple hypothesis testing, and suppose one always uses the most powerful test of level $\alpha = .01$. One can make the frequentist statement that only 1% of true null hypotheses will be rejected (i.e., $R(\theta,\delta) = .01$ for θ equal to the null), but this says nothing about how often one errs when rejecting. For instance, if the test has power of .01 (admittedly terrible power, but useful for making the point) and the null and alternative hypotheses occur equally often in repetitive use of the test, then *half* of all rejections will be in error. Thus one can not make meaningful statements about actual error incurred in repetitive use, without an appropriate bound on $R(\theta,\delta)$ for all θ.

The problem with needing $\bar{R}(\delta)$ is, of course, that it could be a useless bound (or could even be infinite). Indeed, whenever $R(\theta,\delta)$ is highly variable as a function of θ, the reporting of $\bar{R}(\delta)$ is likely to be excessively conservative. The conditional frequentist approaches discussed in Section 2.4 have considerable promise in overcoming this difficulty, however, and can be given interpretations compatible with the Confidence Principle.

Ultimately, the only clear objection to the Confidence Principle is that it conflicts with the LP. This was indicated in the examples and discussion in Chapter 2, and will be seen in later examples also. Most conditionalists view the Confidence Principle, while attractive, as an unattainable goal. (Note, however, that a Bayesian conditionalist follows the Confidence Principle to the extent that his statements of accuracy will be correct, in the long run average sense, if his prior assumptions are correct; one could, indeed, argue that it is the Bayesian who is honestly trying to follow the Confidence Principle by clearly stating the beliefs and assumptions his assessments are based on.) In choosing between the LP and the Confidence Principle, it is important to recall the simple axiomatic basis of the LP, and

to realize that no such basis has been found for the Confidence Principle. Indeed, the long run performance view is deemed rather peculiar by most uninitiated people (c.f., the discussions in the early papers of Neyman in Neyman (1967)).

4.2 THE IRRELEVANCE OF STOPPING RULES

4.2.1 Introduction

One of the most important applications of the LP and RLP is the Stopping Rule Principle (SRP). Stated informally, the SRP is simply that the reason for stopping experimentation (the *stopping rule*) should be irrelevant to evidentiary conclusions about θ. The theoretical and practical implications of the SRP to such fields as sequential analysis and clinical trials are enormous, and will be partially discussed in Sections 4.2.3 and 4.2.4. The SRP itself will be discussed at two levels: in Section 4.2.2 it will be presented in a relatively simple sequential setting, in which it will be shown to follow solely from the LP, while in Section 4.2.6 a very general version will be developed from the RLP. Section 4.2.7 discusses situations in which the SRP is *not* applicable, and Section 4.2.5 points out an interesting conflict between frequentist admissibility and the frequentist belief in the importance of considering stopping rules.

The Stopping Rule Principle was first espoused by Barnard (1947a, 1949), whose motivation at the time was essentially a reluctance to allow an experimenter's *intentions* to affect conclusions drawn from data. (More will be said of this shortly.) The principle was shown to be a consequence of the LP in Birnbaum (1962a), and Barnard, Jenkins and Winsten (1962), and argued to hold in essentially complete generality by Pratt (1965). Other good discussions of the principle can be found in Anscombe (1963), Cornfield (1966), Bartholomew (1967), Basu (1975), Berger (1980), and in many Bayesian works such as Edwards, Lindman, and Savage (1963).

Before formally introducing stopping rules and the stopping rule principle, it is useful to illustrate certain of the ideas through a simple example. The following example, from Berger and Berry (1987), demonstrates the possible extreme dependence of frequentist measures upon the intentions of the experimenter concerning stopping the experiment. The example clearly questions the sensibility of such extreme dependence. (Berger and Berry, 1987, also contains other simple examples, on both sides of the issue.)

EXAMPLE 19.1. A scientist enters the statistician's office with 100 observations, assumed to be independent and from a $N(\theta,1)$ distribution. The scientist wants to test H_0: $\theta = 0$ versus H_1: $\theta \neq 0$. The current average is $\bar{x}_n = 0.2$, so the standardized test statistic is $z = \sqrt{n}|\bar{x}_n - 0| = 2$. A careless classical statistician might simply conclude that there is significant evidence against H_0 at the 0.05 level. But a more careful one will ask the scientist, "Why did you cease experimentation after 100 observations?" If the scientist replies, "I just decided to take a batch of 100 observations," there would seem to be no problem, and very few classical statisticians would pursue the issue. But there is another important question that should be asked (from the classical perspective), namely: "What would you have done had the first 100 observations not yielded significance?"

To see the reasons for this question, suppose the scientist replies: "I would then have taken another batch of 100 observations." This reply does not completely specify a stopping rule, but the scientist might agree that he was implicitly considering a procedure of the form:

(a) take 100 observations;

(b) if $\sqrt{100}|\bar{x}_{100}| \geq k$ then stop and reject H_0;

(c) if $\sqrt{100}|\bar{x}_{100}| < k$ then take another 100 observations and reject if $\sqrt{200}|\bar{x}_{200}| \geq k$.

For this procedure to have level $\alpha = 0.05$, k must be chosen to be 2.18 (Pocock, 1977). Since the actual data had $\sqrt{100}|\bar{x}_{100}| = 2 < 2.18$, the scientist could not actually conclude significance, and hence would have to take the

next 100 observations.

This strikes many people as peculiar. The interpretation of the results of an experiment depends not only on the data obtained and the way it was obtained, but also upon *thoughts* of the experimenter concerning plans for the future.

Of course, this can be carried further. Suppose the puzzled scientist leaves and gets the next 100 observations, and brings them back. Consider two cases. If $\sqrt{200}|\bar{x}_{200}| = 2.1 < 2.18$ then the results are not significant. But they would have been significant had the scientist not paused halfway through the study to calculate z! (It would certainly be tempting not to disclose this interim calculation, and essentially impossible to determine whether or not the scientist had made an interim calculation!) On the other hand, suppose $\sqrt{200}|\bar{x}_{200}| = 2.2 > 2.18$, so now significance has been obtained. But wait! Again the statistician asks what the scientist would have done had the results not been significant. Suppose the scientist says, "If my grant renewal were to be approved, I would then take another 100 observations; if the grant renewal were rejected, I would have no more funds and would have to stop the experiment in any case." The advice of the classical statistician must then be: "We cannot make a conclusion until we find out the outcome of your grant renewal; if it is not renewed, you can claim significant evidence against H_0, while if it is renewed you cannot claim significance and must take another 100 observations." The up-to-now honest scientist has had enough, and he sends in a request to have the grant renewal denied, vowing never again to tell the statistician what he would have done under alternative scenarios.

Note that we are not faulting the classical statistician here for ascertaining and incorporating the stopping rule in the analysis. If one insists on utilization of frequentist measures, such involvement of the stopping rule (even if it exists only in the imagination of the experimenter) is mandatory. The need here for involvement of the stopping rule clearly calls the basic frequentist premise into question, however.

4.2.2 The (Discrete) Stopping Rule Principle

So as not to obscure the essential nature of the SRP, the discussion in this section will be restricted to the following fairly simple situation. Suppose E^τ is a sequential experiment consisting of (i) a sequence of independent observations $X_1, X_2, \ldots$, which will be observed one at a time and which have common density $f_\theta(x)$; and (ii) a non-randomized *stopping rule*, τ, which can be represented by a sequence of sets,

$$A_m \subset \mathcal{X}^m = \mathcal{X}\times\mathcal{X}\times\ldots\times\mathcal{X} \quad \text{(the m-fold Cartesian product of } \mathcal{X}),$$

having the property that

(4.2.1) if $\underset{\sim}{x}^m = (x_1,\ldots,x_m) \in A_m$, sampling stops;

if $\underset{\sim}{x}^m \in A_m^c$, sampling continues.

Since the observations will be observed sequentially, it is clearly unnecessary to have A_m contain points whose first j coordinates were in A_j for any $j < m$; thus we henceforth assume that

$$A_m \cap A_j \times \mathcal{X}^{m-j} = \emptyset \quad \text{for } j < m.$$

The *stopping time*, N, corresponding to τ is that (random) m for which $\underset{\sim}{x}^m \in A_m$; the realization of N will be denoted by n. As usual, only proper stopping rules will be considered, i.e., those which have N finite with probability one for all θ. The probability density of the random experimental outcome $\underset{\sim}{x}^N = (X_1,\ldots,X_N)$ is then

(4.2.2) $$f_\theta^\tau(\underset{\sim}{x}^n) = I_{A_n}(\underset{\sim}{x}^n) \prod_{i=1}^{n} f_\theta(x_i).$$

EXAMPLE 20. Suppose the X_i are $\eta(\theta,1)$.

Case 1. Consider the stopping rule, τ^1, defined by

$$A_m^1 = \begin{cases} \emptyset & \text{if } m \neq k \\ \\ \mathcal{X}^k & \text{if } m = k. \end{cases}$$

The experiment E^{τ^1} is thus the fixed sample size experiment which always observes precisely k observations.

Case 2. Consider the stopping rule, τ^2, defined by

$$A_m^2 = \{\underset{\sim}{x}^m \in \mathcal{X}^m: \ |\bar{x}_m| > Km^{-\frac{1}{2}}\}, \tag{4.2.3}$$

where $\bar{x}_m$ is the mean of $(x_1,\ldots,x_m)$ and K is a fixed positive constant. (By using the Law of the Iterated Logarithm, τ^2 can be shown to be a proper stopping rule.) This stopping rule is rather peculiar, in that it says to stop sampling when the sample mean is K standard deviations from zero.

EXAMPLE 21. Suppose the X_i are Bernoulli (θ).

Case 1. Let E^{τ^1} be the fixed sample size experiment which takes k observations, where $k \geq 2$.

Case 2. Let τ^2 be defined by

$$A_1^2 = \{1\}, \ A_2^2 = \{(0,0),(0,1)\}, \ A_j^2 = \emptyset \quad \text{for } j > 2$$

(i.e., stop if $X_1 = 1$, and otherwise stop after observing X_2), and let E^{τ^2} be the corresponding sequential experiment.

STOPPING RULE PRINCIPLE (SRP): In a sequential experiment E^τ, *with observed final data* $\underset{\sim}{x}^n$, $Ev(E^\tau,\underset{\sim}{x}^n)$ *should not depend on the stopping rule* τ.

The SRP would imply, in Example 20, that if the observation in Case 2 happened to have n = k, then the evidentiary content of the data would be the same as if the data had arisen from the fixed sample size experiment in Case 1. A similar conclusion would hold in Example 21 if n = k occurred.

When $\mathcal{X}$ is discrete, the SRP is an immediate consequence of the LP. This is immediate from (4.2.2) in that $\ell_{\underset{\sim}{x}^n}(\theta)$ is proportional to $\prod_{i=1}^n f_\theta(x_i)$, which does not depend on the stopping rule. For derivation of the SRP in general (from the RLP) see Section 4.2.6.

4.2.3 Positive Implications

A recurring problem in classical statistics is that of optional stopping. Ideally (from a classical viewpoint) an experimenter chooses his stopping rule before experimentation, and then follows it exactly. Actual practice is, however, acknowledged to be quite different. Experiments may end because the data looks convincing enough, because money runs out, or because the experimenter has a dinner date. Indeed, little or no thought may have been given to the stopping rule prior to experimentation, in which case, upon stopping for whatever reason, the data is often treated as having arisen from a fixed sample size design. Optional stopping may often be harmless (such as when the experimenter quits to have dinner), but stopping "when the data looks good" can be a serious error when combined with frequentist measures of evidence. For instance, if one used the stopping rule in Case 2 of Example 20, but analyzed the data as if a *fixed* sample had been taken, one could *guarantee* arbitrarily strong frequentist "significance" against H_0: $\theta = 0$ by merely choosing large enough K.

Optional stopping poses a significant problem for classical statistics, even when the experimenters are extremely scrupulous. Honest frequentists face the problem of getting extremely convincing data too soon (i.e., before their stopping rule says to stop), and then facing the dilemma of honestly finishing the experiment, even though a waste of time or dangerous to subjects, or of stopping the experiment with the prematurely convincing evidence and then not being able to give frequency measures of evidence. One could argue that experiments should be designed to allow for early stopping in response to clear evidence (and, indeed, many such stopping rules have been created, as in the theory of "repeated significance testing"), but there will often be unforeseen eventualities that crop up in sequential experimentation, leaving a strict frequentist in an embarassing position.

Contrast this enormous dilemma with the startling simplicity resulting from use of the SRP. The SRP says that it just doesn't matter; stop for whatever reasons, which (conditional on the data) do not depend on θ (see

Section 4.2.7), and use an appropriate conditional analysis based on $\ell_{\underset{\sim}{x}^n}(\theta)$ (or, alternatively, $\prod_{i=1}^{n} f_\theta(x_i)$). The reason for stopping is simply not relevant. As Edwards, Lindman, and Savage (1963) say

> "The irrelevance of stopping rules to statistical inference restores a simplicity and freedom to experimental design... Many experimenters would like to feel free to collect data until they have either conclusively proved their point, conclusively disproved it, or run out of time, money, or patience."

Anscombe (1963) simply makes the blunt statement "Sequential analysis is a hoax." These comments should be qualified, of course, to the extent that design will depend on the stopping rule. In other words, choosing between two sequential designs obviously involves consideration of stopping rules. Indeed, the most difficult part of (theoretical) sequential (decision) analysis is that of deciding, at a given stage, whether to stop sampling or to take another observation (i.e., choosing the stopping rule). Much of the work done in classical sequential analysis has addressed this problem, and is hence of considerable relevance.

The other desirable implication of the SRP is that analysis of an experiment can be done objectively, in the sense that it is no longer necessary to know the experimenter's intentions towards stopping. It seems very strange that a frequentist could not analyze a given set of data, such as $(x_1,\ldots,x_n)$ in Example 20, if the stopping rule is not given. If the experimenter forgot to record the stopping rule and then died, it is unappealing to have to guess his stopping rule in order to conduct the analysis. As mentioned earlier, it was apparently this feeling, that data should be able to speak for itself, that led Barnard to first support the Stopping Rule Principle.

The above idea is actually a general consequence of the LP, and is useful to apply in areas other than optional stopping. Consider the following example.

EXAMPLE 22. An experiment was conducted with two treatment groups (T_1 and T_2) and a control group (C), the outcomes for each experimental unit being simply success (S) or failure (F). The data was

	C	T_1	T_2
S	8	12	2
F	12	8	8

.

In analyzing the results, the experimenter noted that, in comparing T_1 with C, a standard analysis under the null hypothesis of no treatment effect was not significant at level $\alpha = .1$ (one-tailed), but that if the patients in T_2 and C were pooled, then T_1 was significantly better at the $\alpha = .02$ level. The experimenter went on to say that T_1 was really the treatment of interest and that T_2 was thought to have no effect but was just included for thoroughness, and hence that pooling T_2 and C is acceptable.

To the criticial appraiser, this creates doubts concerning hypothesis selection and confirmation from the same set of data. On the other hand, maybe the experimenter really was planning to pool T_2 and C all along (and was sure T_2 was no worse than C), an especially plausible possibility considering that only 10 patients were given T_2. In any case, it is disconcerting that to analyze the problem from a frequentist perspective we have to know what the experimenter's *intentions* were. Trying to analyze hard data by guessing what the experimenter was thinking before doing the experiment seems rather strange. (Of course, a Bayesian won't necessarily be able to avoid such considerations, since the experimenter's statements may well affect prior probability judgements. The uncertainty will be up front in the prior where it belongs, however, with the data speaking for itself through the likelihood function.)

4.2.4 Criticisms

The rosy statements in the previous section concerning the SRP can be viewed as hopelessly misguided by frequentists, since frequency measures are so dependent on stopping rules. Consider Examples 20 and 21, for instance.

EXAMPLE 21 (continued). In the fixed sample size experiment, $\bar{X}_k$ would be an unbiased estimator of θ for either k = 1 or 2. If one were to ignore the stopping rule, τ^2, in Case 2, however, and still use the sample mean as the estimator, a "problem" of bias arises. Indeed, the sample mean, $\bar{X}_N$, has

$$E_\theta \bar{X}_N = P_\theta(X_1=1)E_\theta[\bar{X}_1|X_1=1] + P_\theta(X_1=0)E_\theta[\bar{X}_2|X_1=0]$$

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

which is biased upwards. Thus if a conditionalist stated he would be using $\bar{X}_N$, regardless of the stopping rule, the experimenter could use τ^2 and "make θ appear larger than it really is" (if desired).

EXAMPLE 20 (continued). This example has been extensively discussed, in terms of its relationship to the SRP and the LP. Armitage (1961) published (to our knowledge) the first such discussion. Basu (1975) gives a particularly thorough examination of a version of the example. For definiteness in highlighting the "paradox," let us assume that a 95% "confidence interval" for θ is desired, and that an "objective" conditionalist states that, if a fixed sample of size n were taken, he would use the interval

$$C_n(\bar{x}_n) = (\bar{x}_n-(1.96)n^{-\frac{1}{2}}, \bar{x}_n + (1.96)n^{-\frac{1}{2}}). \tag{4.2.4}$$

Of course, he would not interpret confidence in the frequency sense, but instead would (probably) use a posterior Bayesian viewpoint with the noninformative prior density $\pi(\theta) = 1$, which leads to a $\eta(\bar{x}_n, n^{-\frac{1}{2}})$ posterior distribution for θ (also, the usual fiducial distribution and the likelihood function for θ).

Suppose now that the experimenter has an interest in seeing that $\theta = 0$ is not in the confidence interval. He could then use the stopping rule

in (4.2.3) for some $K > 1.96$. The conditionalist, being bound to ignore the stopping rule, will still use (4.2.4) as his confidence interval, but this can *never* contain zero. Hence the frequentist probability of coverage of (4.2.4), namely

$$\Gamma(\theta) = P_\theta^{\tau^2}(C_N(\bar{X}_N) \text{ contains } \theta),$$

is such that $\Gamma(0) = 0$ and (by continuity) $\Gamma(\theta)$ is near zero for small θ. The experimenter has thus succeeded in getting the conditionalist to perceive that $\theta \neq 0$, and has done so honestly.

Examples 20 and 21 are typical of how the SRP (or the LP) seems to allow the experimenter to mislead a conditionalist. The "misleading", however, is solely from a frequentist viewpoint, and will not be of concern to a conditionalist. Before discussing why, two comments about Example 20 should be gotten out of the way.

(i) Use of a stopping rule, such as that in (4.2.3), can be chancy for an experimenter if $\theta = 0$ is a real possibility, since then N is likely to be extremely large. (This has no real bearing on the arguments here, however.)

(ii) A Bayesian conditionalist might not completely ignore a stopping rule such as that in (4.2.3), if he suspects it is being used because the *experimenter* thinks θ might be zero. The Bayesian might then assign some positive prior probability, λ, to θ being equal to zero, in recognition of the experimenter's possible knowledge. (The stopping rule is affecting only the prior, however, not "what the data says.") A Bayesian analysis in this situation is strikingly different than that in the "noninformative" case. Indeed, as a particular example, if the $\theta \neq 0$ are given prior density $(1-\lambda)$ times a $\mathcal{N}(0,\rho^2)$ density, then the posterior probability that $\theta = 0$, given the observation $\bar{x}_n = Kn^{-\frac{1}{2}}$, is

$$\pi(0 \mid \bar{x}_n = Kn^{-\frac{1}{2}}) = [1+(\lambda^{-1}-1)(1+n\rho^2)^{-\frac{1}{2}}e^{K^2n\rho^2/2(1+n\rho^2)}]^{-1}.$$

For some specific numbers, suppose that $\rho^2 = 10$, $K = 3$, and $n = 10{,}000$. Then,

$$\pi(0|\bar{x}_n = 3n^{-\frac{1}{2}}) = [1+(\lambda^{-1}-1)(.285)]^{-1}.$$

For moderate λ, this says that $\theta = 0$ is quite plausible when n is large, even though $\bar{x}_n$ is three standard deviations from 0. (This is essentially "Jeffrey's" or "Lindley's" Paradox.)

Finally, let us return to Examples 20 and 21 and see if the conditional perspective might not after all be more intuitively appealing. The use of a biased estimator in Example 21 is really not that troubling, since bias has long been a suspect criterion (especially when compared to, say, the plausibility of the Weak Conditionality Principle). We will concentrate on the more disturbing Example 20, therefore.

EXAMPLE 20 (continued). First of all, the likelihood function for θ (when we stop at time n) is proportional to a $\eta(\bar{x}_n, n^{-\frac{1}{2}})$ density. This clearly indicates that any particular value of θ near $\bar{x}_n$ is more plausible than a value far from $\bar{x}_n$. The interval in (4.2.4) is a reasonable choice from this viewpoint, although other conditionalists might vary the constant 1.96 or shift somewhat towards a suspected prior mean.

Contrast this with the rather unreasonable way in which a frequentist must behave to obtain, say, coverage probability of at least .95 for all θ when K is large. It can be shown that a frequentist should stick to connected intervals (to minimize size for a given coverage probability) and that, when (say) $\bar{x}_n$ is slightly bigger than $Kn^{-\frac{1}{2}}$ and n is fairly large (which will typically be the case for large K and the stopping rule (4.2.3)), these intervals must usually include both zero and $\bar{x}_n$. Hence, in order to ensure the desired coverage probability at zero when K is large, a frequentist will modify (4.2.4) by replacing a small portion of this interval of "likely" θ, such as $(\bar{x}_n + (1.96-\varepsilon_n)n^{-\frac{1}{2}}, \bar{x}_n + (1.96)n^{-\frac{1}{2}})$, with a big interval, $[0, \bar{x}_n-(1.96)n^{-\frac{1}{2}})$, of unlikely θ. This seems unreasonable. The conditionalist knows that an $\bar{x}_n$ satisfying $\bar{x}_n > Kn^{-\frac{1}{2}}$ (with n very large) could have arisen from $\theta = 0$, but

values near $\bar{x}_n$ are so much more likely to be the true θ that he "bets" on these. It should be reemphasized that the conditional analysis is predicated on $\theta = 0$ having no special plausibility; if it does, the Bayesian conclusions (see (ii) above) will be quite different.

The above attempts are probably unlikely to satisfy a frequentist's violated intuition, if the frequentist is not practiced in thinking conditionally. As Savage said in Savage et. al. (1962)

> "I learned the stopping rule principle from Professor Barnard, in conversation in the summer of 1952. Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that some people resist an idea so patently right."

Of some force may be the argument that, if one's intuition gives contradictory insights, it should be trusted in simple situations, such as Example 2, rather than in extremely complex situations such as Example 20. The next section also lends support to the case for ignoring the stopping rule.

4.2.5 Stopping Rules and Inadmissibility

In Section 3.7 it was argued that behavior in violation of the LP, but consistent with the WCP, tends to be decision-theoretically inadmissible. We rephrase the conclusion, in this section, to show that behavior dependent on the stopping rule will often be inadmissible.

Suppose we have possible observations $X_1, X_2, \ldots$, as in Section 4.2.2, and are considering two possible stopping rules, τ^1 and τ^2, with respective stopping sets $\{A_m^1\}$ and $\{A_m^2\}$. The stopping rules, τ^1 and τ^2, are presumed to have the possibility of yielding common data, $\underset{\sim}{x}^n$; i.e., there is presumed to be some n^* and $A \subset A_{n^*}^1 \cap A_{n^*}^2$ such that A has positive probability in both E^{τ^1} and E^{τ^2} for all θ. Examples 20 and 21 are of this type, since the

sets A_k^2 have positive probability for all θ (under both E^{τ^1} and E^{τ^2}), so that $A = A_k^2$ works.

Suppose that we face a decision problem concerning θ, consisting of choice of an action $a \in \mathcal{A}$ under a loss function $L(a,\theta)$ which is strictly convex in "a" for each θ. (More general loss functions can often be handled also.) Proposed for use in E^{τ^1} and E^{τ^2}, respectively, are decision rules $\delta_1(\underset{\sim}{x}^n)$ and $\delta_2(\underset{\sim}{x}^n)$. If, now, the stopping rule is felt to make a difference, δ_1 and δ_2 should differ for at least some of the possible common observations. Thus we suppose that there is some $A^* \subset A$ for which

$$\delta_1(\underset{\sim}{x}^{n^*}) \neq \delta_2(\underset{\sim}{x}^{n^*}) \quad \text{for } \underset{\sim}{x}^{n^*} \in A^*. \tag{4.2.5}$$

Consider, next, the mixed experiment, E^*, consisting of observing $J = 1$ or 2 with probability $\frac{1}{2}$ each and then performing E^{τ^J}. This is a well defined sequential experiment with random observation $(J, \underset{\sim}{X}^{N_J})$, N_J being the stopping time for E^{τ^J}. If the WCP is followed for E^* and (4.2.5) holds, then the decision rule, δ, used for E^* should satisfy

$$\delta((1,\underset{\sim}{x}^{n^*})) \neq \delta((2,\underset{\sim}{x}^{n^*})) \quad \text{for } \underset{\sim}{x}^{n^*} \in A^*.$$

(Alternatively, this inequality should hold on some A^* if it is felt that the stopping rule actually used - i.e., the value of j - really is relevant to the decision.) But, the estimator

$$\delta^*((j,\underset{\sim}{x}^n)) = \begin{cases} \frac{1}{2}\,\delta((1,\underset{\sim}{x}^{n^*})) + \frac{1}{2}\,\delta((2,\underset{\sim}{x}^{n^*})) & \text{if } n = n^* \text{ and } \underset{\sim}{x}^n \in A^* \\ \delta((j,\underset{\sim}{x}^n)) & \text{otherwise} \end{cases}$$

satisfies (because of the strict convexity of L)

$$L(\delta^*((j,\underset{\sim}{x}^{n^*})),\theta) < \tfrac{1}{2}\,L(\delta((1,\underset{\sim}{x}^{n^*})),\theta) + \tfrac{1}{2}\,L(\delta((2,\underset{\sim}{x}^{n^*})),\theta). \tag{4.2.6}$$

Hence, letting E_θ^*, $E_\theta^{\tau^1}$, and $E_\theta^{\tau^2}$ stand for expectation in experiments E^*, E^{τ^1}, and E^{τ^2}, respectively, the frequentist risk (in E^*) of δ^* satisfies

$$
\begin{aligned}
R(\theta,\delta^*) &= E^*_\theta L(\delta^*((J,\underset{\sim}{X}^{N_J})),\theta) \\
&= \tfrac{1}{2} E^1_\theta L(\delta^*((1,\underset{\sim}{X}^{N_1})),\theta) + \tfrac{1}{2} E^2_\theta L(\delta^*((2,\underset{\sim}{X}^{N_2})),\theta) \\
&< \tfrac{1}{2} E^1_\theta L(\delta((1,\underset{\sim}{X}^{N_1})),\theta) + \tfrac{1}{2} E^2_\theta L(\delta((2,\underset{\sim}{X}^{N_2})),\theta) \\
&= E^*_\theta L(\delta((J,\underset{\sim}{X}^{N_J})),\theta) \\
&= R(\theta,\delta).
\end{aligned}
$$

(The inequality above is strict because of (4.2.6), the fact that A* has positive probability for all θ in E^{τ^1} and E^{τ^2}, and providing $R(\theta,\delta)$ is finite.) This establishes the inadmissibility of allowing the stopping rule to affect the decision making.

EXAMPLE 21 (continued). Suppose that the goal is to estimate θ under squared error loss, and that, because of the bias in use of $\bar{X}_N$ for the stopping rule τ^2, an estimator $\delta_2(\underset{\sim}{x}^n)$ would be used (in E^{τ^2}) such that $\delta_2(\underset{\sim}{x}^n)$ is *not* equal to $\bar{x}_n$ for at least one possible observation, say, $n = 1$, $x_1 = 1$. Let E^{τ^1} be the fixed sample size experiment of size $k = 1$, and suppose that $\delta_1(x_1) = x_1$ would be used for this experiment. However, the experimenter chooses between performing E^{τ^1} and E^{τ^2} on the basis of a fair coin flip ($J = 1$ or 2). This is exactly the situation discussed above, and if the experimenter follows his "instincts" and uses different estimates (depending on J or the actual stopping rule employed)when $x_1 = 1$ is observed, he will be behaving in an inadmissible fashion.

The development above is just a special case of that in Section 3.7, which in turn is basically just a version of the Rao-Blackwell theorem. (Here, J is *not* part of the sufficient statistic for θ in E* when $\underset{\sim}{x}^{n^*} \in A^*$, and decision rules should be based only on the sufficient statistic.) The reason for explicitly giving the development in the sequential framework is to clearly exhibit the conflict between the frequentist desire for admissibility and the intuitive notion that the stopping rule used should matter.

4.2.6 The General Stopping Rule Principle

The SRP can be generalized to an essentially arbitrary sequence of experiments, and shown (in this generality) to follow from the RLP. Thus suppose we have available a sequence $E_1, E_2, \ldots$ of experiments (replacing the i.i.d. observations, $X_1, X_2, \ldots$, of Section 4.2.2) consisting of observing X_j on $\mathcal{X}_j$. We can consider, for each m, the composite experiment $E^m = (\underset{\sim}{\mathcal{X}}^m, \theta, \{P_\theta^m\})$ consisting of observing $\underset{\sim}{X}^m = (X_1, \ldots, X_m)$ on $\mathcal{X}^m = \prod_{j=1}^m \mathcal{X}_j$ with probability distribution P_θ^m. (If the experiments are independent, P_θ^m will simply by the product measure of the individual distributions on $\mathcal{X}_j$.)

We consider sequential procedures in which we decide, after performing experiments $E_1, \ldots, E_m$, whether or not to perform E_{m+1}. As usual, we can allow this decision to depend upon the outcome of an auxilliary chance mechanism, leading to the following general notion of a stopping rule.

DEFINITION. *A <u>stopping rule</u> is a sequence* $\underset{\sim}{\tau} = (\tau_0, \tau_1, \ldots)$ *in which* $\tau_0 \in [0,1]$ *is a constant and* τ_m: $\mathcal{X}^m \to [0,1]$ *a measurable function on* $\mathcal{X}^m$ *for* $m \geq 1$.

The intention is that $\tau_m(\underset{\sim}{x}^m)$ represent the conditional probability of stopping after only m observations, given that we have taken m observations and have observed $\underset{\sim}{x}^m = (x_1, \ldots, x_m)$. The nonrandomized stopping rules discussed in Section 4.2.2 are the special case where the τ_m can only assume the values 0 and 1. When convenient, we shall regard τ_0 as a function on the one-point set $\mathcal{X}^0 = \{\emptyset\}$, the "sample space" for the null experiment $E^0 = (\mathcal{X}^0, \theta, \{P_\theta^0\})$, with P_θ^0 the point mass at $\mathcal{X}^0$'s only point for all θ.

Now define $\mathcal{X}^* = \{(m, \underset{\sim}{x}^m): m \in \mathbb{N}, \underset{\sim}{x}^m \in \mathcal{X}^m\}$. For $\underset{\sim}{x}^m = (x_1, \ldots, x_m) \in \mathcal{X}^m$ and $0 \leq j \leq m$, let $\underset{\sim}{x}^{m,j} = (x_1, \ldots, x_j) \in \mathcal{X}^j$ denote the initial segment; of course $\underset{\sim}{x}^0 = \emptyset \in \mathcal{X}^0$ no matter what $\underset{\sim}{x}^m \in \mathcal{X}^m$ might be. For each stopping rule, $\underset{\sim}{\tau}$, determine a family $\{P_\theta^{\underset{\sim}{\tau}}\}$ of measures on $\mathcal{X}^*$ by setting

$$P_\theta^{\underset{\sim}{\tau}}(m,A) = \int_A \prod_{j=0}^{m-1} (1-\tau_j(\underset{\sim}{x}^{m,j}))\tau_m(\underset{\sim}{x}^m) P_\theta^m(d\underset{\sim}{x}^m)$$

for each m and Borel set $A \subset \mathcal{X}^m$. With this definition, τ_0 is the probability of performing E^0, i.e. of taking no data at all. After observing $\underset{\sim}{x}^m$, $\tau_m(\underset{\sim}{x}^m)$ is the conditional probability of taking no more observations.

If $P_\theta^{\underset{\sim}{\tau}}(\mathcal{X}^*) = 1$ for all θ, then the procedure is certain to stop eventually and $\underset{\sim}{\tau}$ is called *proper*; otherwise $\underset{\sim}{\tau}$ is improper and, for at least one θ, there is a positive probability $(1-P_\theta^{\underset{\sim}{\tau}}(\mathcal{X}^*))$ that the sequential procedure would require sampling an infinite number of times. For a proper stopping rule, τ, we can consider the sequential experiment

$$E^{\underset{\sim}{\tau}} = ((N,\underset{\sim}{X}^N),\ \theta,\ \{P_\theta^{\underset{\sim}{\tau}}\}),$$

where N denotes the (random) stopping time. (It is notationally convenient to include N as part of the observation although it could, of course, be recovered from $\underset{\sim}{x}^N$.)

The Stopping Rule Principle for this general setting is formalized in the following theorem, and is shown to follow from the RLP.

THEOREM 5 (The Stopping Rule Principle). *From the* RLP, *it follows that, for any (proper) stopping rule* $\underset{\sim}{\tau}$,

$$Ev(E^{\underset{\sim}{\tau}},(n,\underset{\sim}{x}^n)) = Ev(\underset{\sim}{E}^n,\underset{\sim}{x}^n)$$

for $\{P_\theta^{\underset{\sim}{\tau}}\}$-a.e. $(n,\underset{\sim}{x}^n)$, *i.e. the evidence concerning* θ *in* $E^{\underset{\sim}{\tau}}$ *is identical with that for the fixed sample size experiment* $\underset{\sim}{E}^n$ *(with the observed* n*), so that* $\underset{\sim}{\tau}$ *is irrelevant.*

Proof. Pick $n \in \mathbb{N}$ and let $U_1 \subset \mathcal{X}^*$ be the set of points $(n,\underset{\sim}{x}^n)$ with $\underset{\sim}{x}^n \in \mathcal{X}^n$ satisfying $0 < \tau_n(\underset{\sim}{x}^n)\prod_{j=0}^{n-1}(1-\tau_j(\underset{\sim}{x}^{n,j}))$, and let $c\colon\ U_1 \to (0,\infty)$ be the indicated product. Map U_1 onto $U_2 = \{\underset{\sim}{x}^n \in \mathcal{X}^n\colon\ (n,\underset{\sim}{x}^n) \in U_1\}$ by setting $\varphi(n,\underset{\sim}{x}^n) = \underset{\sim}{x}^n$. Then φ is one-to-one and bimeasurable, and

$$P_\theta^n(A) = \int_{\varphi^{-1}(A)} [1/c(\underset{\sim}{y})]P_\theta^{\underset{\sim}{\tau}}(d\underset{\sim}{y}).$$

The assertion of the theorem now follows from the RLP. ||

Notice that Θ was not required to be a subset of some Euclidean space, nor was $\{P_\theta^m\}$ required to be a dominated family; thus even in situations where no version of the usual LP can apply, the SRP is valid (provided, of course, that the WCP and SP, and hence the RLP, are accepted). This was observed in Pratt (1965).

4.2.7 Informative Stopping Rules

Even the definition of a stopping rule given in the last section may seem somewhat narrow when compared with the vast possibilities for informal stopping discussed in Section 4.2.3. Stopping rules which appear to be more general can be created by introducing an auxilliary variable Y (possibly random), and allowing τ_m, the conditional probability of stopping at stage m, to depend on the value of Y, as well as on $\underset{\sim}{x}^m$. This actually adds very little generality, however, since the values of Y at each stage could simply be incorporated into the data X_i. The following example illustrates the importance of sometimes doing this.

EXAMPLE 23. Suppose $X_1, X_2, \ldots$ are independent Bernoulli (θ) random variables, with $\theta = .49$ or $\theta = .51$. The observations, however, arrive randomly. Indeed, if $\theta = .49$, the observations arrive as a Poisson process with mean rate of 1 per second, while if $\theta = .51$, the observations will arrive as a Poisson process with mean rate of 1 per hour. The "stopping rule" that will be used is to stop the experiment at the first observation that arrives *after* 1 minute has elapsed. One can here introduce Y = time, and write down the stopping rule in terms of Y and the X_i.

It is intuitively clear that this stopping rule cannot be ignored since, if one ends up with 60 observations, knowing whether the experiment ran for 1 minute or 2 $\frac{1}{2}$ days is crucial knowledge. Incorporating Y into the data resolves all ambiguities, however. Thus, simply define Y_i as the (random) time at which the $i^{\underline{th}}$ observation arrives, and consider the experiment to consist of observing (X_1,Y_1), $(X_2,Y_2), \ldots$. The stopping rule will be given by

$$\tau_m(((x_1,y_1),\ldots,(x_m,y_m))) = \begin{cases} 0 & \text{if } y_m < 1 \\ \\ 1 & \text{if } y_m \geq 1, \end{cases}$$

and is of the form discussed in Section 4.2.6 (or even Section 4.2.2). The importance of the number of observations arriving during the time span of the experiment will be reflected in the portion of the likelihood function due to the y_i.

Slightly more generality might be needed than afforded by simply observing the auxilliary variables at the observation times (as in Example 23) and including them as part of the observations, but the idea is clear: consider *all* available observational information as part of the data X_i. (Of course, some auxilliary information may be considered too informal to include as part of the data, and yet may have some effect on stopping, but such information should only be ignored if it seems relatively unimportant, in which case its effect on stopping can probably also be ignored.)

Even within the above more general perspective on stopping rules, a difficulty might still arise. This difficulty is that the stopping rule might be unknown or partially unknown, in that cessation of the sequential experiment could depend on unobservable random quantities whose probability distributions are not completely known. Following the convention of Section 3.5 and letting θ denote *all* unknown quantities, we could thus write a general stopping rule in terms of $\tau_m(\underset{\sim}{x}^m,\theta)$. (Actually, by including a uniform random variable in θ, it would be possible to have the τ_m assume only the values zero or one.) The general density on $\mathcal{X}^*$ (densities, and discreteness if necessary, being assumed to retain compatibility with Section 3.5) would then be

$$f^*_\theta((n,\underset{\sim}{x}^n)) = [\prod_{j=0}^{n-1}(1-\tau_j(\underset{\sim}{x}^n,j,\theta))]\tau_n(\underset{\sim}{x}^n,\theta)f^n_\theta(\underset{\sim}{x}^n),$$

where f^n_θ is the density corresponding to P^n_θ. Again following Section 3.5, one

could write $\theta = (\xi,\eta)$, where ξ is of interest and η is a nuisance variable. If, for the observed $(n,\underset{\sim}{x}^n)$, $\tau_j(\underset{\sim}{x}^{n,j},\theta)$ does not depend on ξ for $j \leq n$, and if η is a noninformative nuisance parameter (see Section 3.5) for the fixed sample size experiments involving observation of $\underset{\sim}{X}^n$, then the LP and NNPP (see Section 3.5) imply that $\underset{\sim}{\tau}$ is irrelevant. Such a $\underset{\sim}{\tau}$ is called *noninformative*; otherwise $\underset{\sim}{\tau}$ is said to be *informative* and the SRP will not apply. (Raiffa and Schlaifer (1961) introduced these terms.)

We do not pursue the matter further because informative stopping rules occur only rarely in practice (providing all observational information is incorporated into the X_i, as in Example 23). There exists a certain amount of disagreement concerning this point, but the disagreement seems to be primarily due to the misconception that an informative stopping rule is one for which N carries information about θ. This is *not* the definition of an informative stopping rule. Very often N *will* carry information about θ, but to be informative a stopping rule must carry information about θ additional to that available in $\underset{\sim}{X}^N$, and this last will be rare in practice.

4.3 THE IRRELEVANCE OF CENSORING MECHANISMS

4.3.1 Introduction

Another great simplification that application of the LP (or RLP) makes possible is in the handling of censoring. Data is often observed in censored form, and the mechanisms causing the censoring can be quite involved. In most such cases, the LP (or RLP) will imply that only the result of the censoring, and not the censoring mechanism or distribution, is relevant to conclusions about θ.

Section 4.3.2 considers the situation of fixed (nonrandom) censoring, and establishes a version of the irrelevance of censoring mechanisms called Censoring Principle 1. One of the implications of Censoring Principle 1 is that the evidential import of an uncensored observation, from an experiment in which censoring was possible, is the same as the identical observation from an uncensored version of the experiment.

Section 4.3.3 considers random censoring, and establishes conditions under which the distribution of the censoring random variable is irrelevant. The main condition is on the censoring mechanism itself, and leads to the concept of a *noninformative censoring mechanism*. This concept is surprisingly simple and powerful. It is not the case, however, that all sensible censoring mechanisms are noninformative, although many common ones are. This issue is discussed in Section 4.3.4.

The Censoring Principle, as it applies to uncensored observations in nonrandom censoring, seems to be due to John Pratt (see Pratt (1961, 1965), his discussion in Birnbaum (1962a), and the discussion in Savage et. al. (1962)). The general Censoring Principles developed here and the concept of a noninformative censoring mechanism appear to be new, however. Before proceeding with these general developments, it is worthwhile to present an illuminating (and entertaining) example from Pratt's discussion of Birnbaum (1962a). The example makes a simple version of the Censoring Principle seem intuitively obvious.

EXAMPLE 24 (Pratt). A sample of 25 observations was taken from a $\eta(\theta,\sigma^2)$ population, and inference about the population mean was desired. All observations were found to lie between 72 and 99, and a standard normal analysis was performed by a frequentist statistician. The statistician reported the analysis to the experimenter, but, curious about the observed 99, asked the experimenter how high his measuring instrument (assumed to be perfectly accurate) read. The experimenter said that the instrument only read to 100, but that, if he had observed a reading of 100, he would have switched to another instrument which had a range of 100 to 1000. The statistician was happy with this response, and satisfied with a job well done.

The next day the experimenter called about something else, and mentioned that he had just checked the high range instrument and found that it was broken. The statistician asked if the experimenter would have had the instrument repaired before completing the previous experiment, to which the

experimenter said no. The statistician then said that what were really being observed were observations, X_i, from the truncated distribution with the usual normal density for $x_i < 100$ but the point mass

$$P_{\theta,\sigma^2}(100) = \int_{100}^{\infty} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp\{-\frac{1}{2\sigma^2}(x-\theta)^2\}dx$$

at $x_i = 100$. This, said the statistician, calls for a different analysis; for instance, the usual $100(1-\alpha)\%$ confidence interval for θ in the normal situation would no longer have probability of coverage of at least $1-\alpha$ in the truncated situation. The experimenter reacted to this with outrage, saying that he observed precisely what he would have observed had the high range instrument been working (all observations *were* less than 100), and that the condition of an instrument never used in the experiment hardly seemed relevant to the information about θ obtained from the experiment. The frequentist statistician merely shook his head at the naivete of experimenters.

4.3.2 Fixed Censoring and Equivalent Censoring Mechanisms

Consider an experiment $E = (X, \theta, \{P_\theta\})$. Fixed censoring occurs when, instead of X, one observes $Y = g(X)$, where g is a known function from $\mathcal{X}$ into $\mathcal{Y}$. Thus the experiment really performed is $E^g = (Y, \theta, \{P_\theta o g^{-1}\})$. (As usual, if $A \subset \mathcal{Y}$, $g^{-1}(A) = \{x \in \mathcal{X}: g(x) \in A\}$.)

EXAMPLE 25. Suppose $X = (X_1,\ldots,X_n)$, where the X_i represent the times of death due to cancer of patients in a cancer survival experiment. Suppose, however, that the experiment will last only ten years, so that the real data will, for the $i^{\underline{th}}$ patient, be

$$Y_i = (Y_i^1, Y_i^2) \equiv (\min\{X_i, 10\}, I_{[0,10]}(X_i)) \tag{4.3.1}$$

(i.e., the truncated survival time and an indicator as to whether the observation is or is not truncated). Thus

$$Y = g(X) \equiv (Y_1,\ldots,Y_n). \tag{4.3.2}$$

This is an example of what is commonly called type I censoring. Example 24 is

also of this type.

EXAMPLE 26. Suppose that $X = (X_1,\ldots,X_n)$, but that the n-r largest of the X_i will be truncated at the $r^{\underline{th}}$ largest. Thus let

$$Y_i = (Y_i^1, Y_i^2) \equiv (\min\{X_i, X_{(r)}\},\ I_{[-\infty, X_{(r)}]}(X_i)), \tag{4.3.3}$$

where $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ are the order statistics for X. Again

$$Y = g(X) \equiv (Y_1,\ldots,Y_n). \tag{4.3.4}$$

This is an example of what is commonly called type II censoring.

EXAMPLE 27. Suppose $\mathcal{X} = R^n$, $\mathcal{Y} = \mathcal{X} \times \{0,1\}$, and for some fixed $\rho > 0$,

$$g(X) = \begin{cases} (X,0) & \text{if } |X| \leq \rho \\ (\rho X/|X|, 1) & \text{if } |X| > \rho. \end{cases} \tag{4.3.5}$$

Then E^g represents the experiment in which the radius of X is truncated at ρ, but the direction, $X/|X|$, of X is faithfully reported. This is not a standard "type" of censoring, but fits easily within our framework.

Our goal in this section is to indicate that the only effect a censoring mechanism should have on a conclusion is to convey knowledge concerning the actual location of x in $\mathcal{X}$. This may seem intuitively obvious, but Example 24 is a prime illustration of how this is not the case classically. We formalize this notion in the following definition.

DEFINITION. *Let* $E = (X, \theta, \{P_\theta\})$ *be a given uncensored experiment, and consider two fixed censoring mechanisms* g_1 *and* g_2. *These mechanisms will be said to be* <u>equivalent on A</u> $\subset \mathcal{X}$ *if, for all* $x \in A$,

$$g_1^{-1}(g_1(x)) = g_2^{-1}(g_2(x)). \tag{4.3.6}$$

As a special case, a single fixed censoring mechanism, g, *will be said to be* <u>equivalent to no censoring on A</u> $\subset \mathcal{X}$ *if* $g^{-1}(g(x)) = x$ *for all* $x \in A$.

The idea in the above definition is that, for censoring mechanism g_i, one observes $Y_i = g_i(X)$ and that the only information communicated by the censored data, y_i, is that x was in $g_i^{-1}(y_i)$. If (4.3.6) is satisfied, then g_1 and g_2 will always convey the same information (for $x \in A$). And a g which is equivalent to no censoring (for $x \in A$) conveys exactly the same information that x does. In Example 24, it is clear that the censoring mechanism is equivalent to no censoring on $A = \{x: \ x_i < 100, \ i = 1,\ldots,25\}$; in Example 25, g is equivalent to no censoring on $A = \{x: \ x_i < 10, \ i = 1,\ldots,n\}$; and, in Example 27, g is equivalent to no censoring on $A = \{x: \ |x| < \rho\}$. As an example of possible equivalence of two different censoring mechanisms, consider the following combination of Examples 25 and 26.

EXAMPLE 28. Suppose $X = (X_1,\ldots,X_n)$, where the X_i can assume only positive integer values. Let g_1 be as in (4.3.1) and (4.3.2), g_2 be as in (4.3.3) and (4.3.4), and $A = \{x: \ x_{(r)} = 10\}$. It is easy to check that, for $x \in A$,

$$g_1^{-1}(g_1(x)) = g_2^{-1}(g_2(x)) = \{z \in A: \ z_i = x_i \ \text{ if } x_i \leq 10\}.$$

Hence the type I and type II censoring would, in this case, be equivalent on A. (Note that classical analysis tends to treat the two types of censoring differently.)

We now formally state, and justify, the principle that equivalent censoring mechanisms convey the same information about θ, for $x \in A$.

CENSORING PRINCIPLE 1. *If* E^{g_1} *and* E^{g_2} *are two experiments arising from censoring mechanisms equivalent on* A *for an experiment* E, *then*

$$\text{Ev}(E^{g_1}, g_1(x)) = \text{Ev}(E^{g_2}, g_2(x)) \tag{4.3.7}$$

for all $x \in A$ *if* $\mathcal{X}$ *is discrete, and for* $\{P_\theta\}$ - a.e. $x \in A$ *in general. As a special case, if* $g^{-1}(g(x)) = x$ *for all* $x \in A$, *then* (4.3.7) *can be replaced by*

$$\text{Ev}(E^{g}, g(x)) = \text{Ev}(E,x). \tag{4.3.8}$$

Censoring Principle 1 follows from the LP in the discrete case

since, by definition, the probabilities of $g_1(x)$ and $g_2(x)$ are equal (to $P_\theta(g_i^{-1}(g_i(x)))$ for all θ. In the general case it follows from the RLP by setting $U_1 = \{g_1(x): x \in A\}$, $U_2 = \{g_2(x): x \in A\}$, $\varphi(g_1(x)) = g_2(x)$ and $c(g_1(x)) = 1$ for $x \in A$.

The greatest practical use of Censoring Principle 1 is in the case where a censoring mechanism, g, is equivalent to no censoring on A, as was the case in Examples 24, 25, and 27 when no censoring happened to occur. The censoring mechanism can then be completely ignored.

4.3.3 Random Censoring

To generalize the notion of censoring to include random censoring, let $\lambda \in \Lambda$ be a censoring variable with probability density ν on Λ. (To avoid technicalities, discreteness of Λ and $\mathcal{X}$ will be assumed until the end of the section.) Suppose that X and λ are independent (without which very little progress can be made), and that

$$Y = g(X,\lambda) \in \mathcal{Y}$$

is observed. The actual experiment performed can thus be written

$$E^{g,\nu} = (Y, \theta, \{f_\theta^{g,\nu}\}),$$

where the density of Y is

$$f_\theta^{g,\nu}(y) = \sum_{\{(x,\lambda):\ g(x,\lambda) = y\}} f_\theta(x)\nu(\lambda) \quad . \tag{4.3.9}$$

EXAMPLE 29. Suppose X represents the time at which a patient in a cancer survival study would suffer death due to cancer, and let λ represent the death time due to competing risks. (We will sidestep the issue of whether or not X and λ can be well-defined.) The actual observation for the patient will be

$$Y = (Y^1, Y^2) = g(X,\lambda) \equiv (\min\{X,\lambda\},\ I_{[0,\lambda]}(X)), \tag{4.3.10}$$

i.e., the actual time of death, Y^1, and an indicator, Y^2, as to the cause of death. Generalization to involve data from n patients and a variety of

competing risks is straightforward, and all the subsequent theory will apply equally well to such a generalization.

The LP, of course, implies that the likelihood function, determined from (4.3.9) for the observed y, contains all the information about θ available from the experiment. The difficulty in utilizing this likelihood function lies in the presence of ν in the expression: typically, ν will be unknown (and complicated). If, however, ν were judged to convey no information about θ (see Section 3.5 and Section 4.3.4) *and* $f_\theta^{g,\nu}(y)$ could be shown to factor into separate terms involving θ and ν, then the difficulty would disappear. This would result in an enormous simplification of the analysis, and is another of the great practical gains that can be realized through adoption of the LP. The following definition gives the key characterization of censoring mechanisms for which this program is possible.

DEFINITION. *A censoring mechanism* $g: \mathcal{X} \times \Lambda \to \mathcal{Y}$ *is* <u>noninformative</u> *at* $y \in \mathcal{Y}$ *if* $g^{-1}(y)$ *is a product set, i.e., if*

$$g^{-1}(y) = A_y \times B_y, \text{ where } A_y \subset \mathcal{X} \text{ and } B_y \subset \Lambda.$$

EXAMPLE 29 (continued). Here

$$g^{-1}((y^1,y^2)) = \begin{cases} (y^1,\infty) \times \{y^1\} & \text{if } y^2 = 0 \\ \{y^1\} \times [y^1,\infty) & \text{if } y^2 = 1, \end{cases}$$

so that g is a noninformative censoring mechanism at *all* $y \in \mathcal{Y}$.

EXAMPLE 27 (continued). Consider the situation in Example 27, but assume that ρ is now a random variable (and, hence, replace $g(X)$ by $g(X,\rho)$). Since

$$g^{-1}((y^1,y^2)) = \begin{cases} \{y^1\} \times [|y^1|,\infty) & \text{if } y^2 = 0 \\ \{cy^1\colon\ c > 1\} \times \{|y^1|\} & \text{if } y^2 = 1, \end{cases}$$

g is a noninformative censoring mechanism at *all* $y \in \mathcal{Y}$.

If g is noninformative at y, then (4.3.9) becomes (employing also the independence of X and λ)

$$f_\theta^{g,\nu}(y) = [\sum_{x\in A_y} f_\theta(x)][\sum_{\lambda\in B_y} \nu(\lambda)], \tag{4.3.11}$$

so that (for known ν), the LP implies that all information concerning θ from the experiment is contained in

$$\ell_y^*(\theta) = \sum_{x\in A_y} f_\theta(x). \tag{4.3.12}$$

If ν is unknown but "noninformative" for θ (see Sections 3.5 and 4.3.4), the same conclusion follows from the NNPP in Section 3.5.5. These conclusions can be summarized as follows.

CENSORING PRINCIPLE 2. *If* $\mathcal{X}$ *and* Λ *are discrete,* X *and* λ *are independent,* g *is noninformative at the observed* y, *and either* ν *is known or it is unknown but noninformative, then* $\mathrm{Ev}(E^{g,\nu},y)$ *depends only on* $\ell_y^*(\theta)$ (*from* (4.3.12)).

Note that this principle does *not* say that censoring has no effect on the analysis. Indeed, $\ell_y^*(\theta)$ will often fail to be proportional to $\ell_x(\theta) = f_\theta(x)$, which would be used if no censoring occurred. Another point is that the only censoring mechanisms which can *guarantee* that $\mathrm{Ev}(E^{g,\nu},y)$ does not depend on ν (for ν as in the principle) are noninformative censoring mechanisms. This is established in the following theorem.

THEOREM 6. *If* g: $\mathcal{X} \times \Lambda \to \mathcal{Y}$ *is* <u>*not*</u> *a noninformative censoring mechanism at* y, *then there exists* $\{f_\theta\}$ *on* $\mathcal{X}$ *such that* $\mathrm{Ev}(E^{g,\nu},y)$ *depends on* ν.

Proof. If $g^{-1}(y)$ is not a product set, it follows that there exist two points $\lambda_1, \lambda_2 \in \Lambda$ such that either

$$\Omega_1 = \{x:\ g(x,\lambda_1) = y \text{ and } g(x,\lambda_2) \neq y\},$$

or

$$\Omega_2 = \{x:\ g(x,\lambda_1) \neq y \text{ and } g(x,\lambda_2) = y\},$$

or both are nonempty. Consider ν that are concentrated on $\{\lambda_1,\lambda_2\}$, and define

$$\Omega_3 = \{x: \ g(x,\lambda_1) = g(x,\lambda_2) = y\}.$$

Equation (4.3.9) can then be written

$$f_\theta^{g,\nu}(y) = \nu(\lambda_1)P_\theta(\Omega_1) + \nu(\lambda_2)P_\theta(\Omega_2) + P_\theta(\Omega_3).$$

$$= \nu(\lambda_1)[P_\theta(\Omega_1) - P_\theta(\Omega_2)] + P_\theta(\Omega_2 \cup \Omega_3).$$

Thus, as long as $\{f_\theta\}$ is chosen so that $[P_\theta(\Omega_1)-P_\theta(\Omega_2)]$ and $P_\theta(\Omega_2 \cup \Omega_3)$ are not proportional as functions of θ, the likelihood function will depend on $\nu(\lambda_1)$. ||

Finally, we leave the discrete setting and develop a very general version of Censoring Principle 2, based on the RLP. We will assume that Λ and $\mathcal{Y}$ are LCCB spaces, that ν is a Borel probability measure, and that $g: \mathcal{X} \times \Lambda \to \mathcal{Y}$ is a Borel function. The actual experiment of observing $Y = g(X,\lambda)$ is $E^{g,\nu} = (Y, \theta, \{P_\theta^{g,\nu}\})$, where

$$P_\theta^{g,\nu}(C) = (P_\theta \times \nu)(\{(x,\lambda): \ g(x,\lambda) \in C\}). \tag{4.3.13}$$

The definition of a noninformative censoring mechanism at y remains unchanged, and leads to the following principle.

CENSORING PRINCIPLE 2'. *Let* $C \subset \mathcal{Y}$ *be a Borel set such that* g *is a noninformative censoring mechanism at all* $y \in C$. *Suppose* ν_1 *and* ν_2 *are Borel probability measures (for* λ*) which are mutually absolutely continuous on* $C^* = \bigcup_{y \in C} B_y$ (*where* $g^{-1}(y) = A_y \times B_y$). *Then, if either* (i) ν_1 *and* ν_2 *are known, or* (ii) *they are unknown but noninformative (see Sections* 3.5 *and* 4.3.4*), it should be the case that*

$$Ev(E^{g,\nu_1},y) = Ev(E^{g,\nu_2},y) \quad \textit{for } \{P_\theta^{g,\nu_1}\}\text{-a.e. } y \in C. \tag{4.3.14}$$

The conclusion in Censoring Principle 2' is not quite as strong as that in the original Censoring Principle 2, in that evidentiary equivalence is only stated to hold among equivalence classes of ν (on C). Of course, if the possible ν under consideration are known to be absolutely continuous with respect to some measure μ, then it can be stated that ν is irrelevant (if it is

noninformative). For instance, in Example 29 it may be reasonable to assume that ν is absolutely continuous with respect to Lebesgue measure, and is thus ignorable (if noninformative).

It seems likely that Censoring Principle 2' is a general consequence of the RLP. This is because one can define (see the RLP) $U_1 = U_2 = C$, φ to be the identity map, and

$$c(y) = c((x,\lambda)) = \nu_2(d\lambda)/\nu_1(d\lambda), \tag{4.3.15}$$

and seek to show that (for any Borel subset, D, of C)

$$P_\theta^{g,\nu_2}(D) = \int_D c(y) P_\theta^{g,\nu_1}(dy). \tag{4.3.16}$$

Since (4.3.16) is essentially (3.4.1) of the RLP (where 1/c has been replaced by c for convenience in what follows), Censoring Principle 2' would be an immediate consequence of the RLP (and the NNPP of Section 3.5, if the ν_i are unknown but noninformative). And (4.3.16) seems to be a correct equation: it can trivially be verified to hold in the discrete setting, for instance. Unfortunately, severe measurability difficulties (due to the possible nasty nature of g) prevented us from verifying (4.3.16), in general. Under additional conditions, however, we were able to show that (4.3.16) does hold for some positive c, which suffices, by the above argument, to establish Censoring Principle 2' as a consequence of the RLP. Furthermore, though somewhat technical, these additional conditions involve only the censoring mechanism, g, and not the P_θ or ν. This makes general verification of the irrelevance of any specific censoring mechanism possible.

THEOREM 7. *Let* g *be a noninformative censoring mechanism at all* $y \in C$, *and suppose that there exist sequences* $\{\varphi_n\}$ *and* $\{\psi_n\}$ *of measurable mappings* φ_n: $\mathcal{X} \to \mathcal{X}$ *and* ψ_n: $\Lambda \to \Lambda$, *such that the functions* $g_n(x,\lambda) \equiv g(\varphi_n(x),\psi_n(\lambda))$ *are countably valued and the* σ*-algebras,* $\mathcal{L}$, $\mathcal{F}_n$, *and* $\mathcal{L}_n$, *generated on* $\mathcal{X} \times \Lambda$ *by* $g(x,\lambda)$, $\varphi_n(x)$, *and* $g_n(x,\lambda)$, *respectively, satisfy the conditions*

i) $$\mathcal{F}_n \vee \mathcal{L}_n \subset \mathcal{F}_{n+1} \vee \mathcal{L}_{n+1}$$

ii) $$\bigcap_{m=1}^{\infty} \bigvee_{n=m}^{\infty} \mathcal{B}_n = \mathcal{B}.$$

Then for any two probability measures ν_1 *and* ν_2 *on* Λ, *which are mutually absolutely continuous on* C,

$$\int_C h(y) P^{g,\nu_2}(dy) = \int_C h(y)c(y) P^{g,\nu_1}(dy) \tag{4.3.17}$$

for every bounded measurable function h *on* $\mathcal{Y}$ *and every probability measure* P *on* $\mathcal{X}$. *(Note that* (4.3.16) *follows trivially from* (4.3.17). *Hence, under the above conditions, Censoring Principle* 2' *is a consequence of the* RLP.*)*

Proof. We will prove the theorem for $C = \mathcal{Y}$. The modifications needed for arbitrary C are obvious. For $n \geq 1$ let $\{y_j^n\}_{j\geq 1}$ be the countably many values of g_n; the σ-algebra $\mathcal{B}_n$ is generated by the countable partition $\mathcal{P}^n = \{A_j^n \times B_j^n\}$ of $\mathcal{X} \times \Lambda$ into the measurable rectangles (or product sets) $A_j^n \times B_j^n = g_n^{-1}(y_j^n)$, where $A_j^n = \varphi_n^{-1}(A_{y_j^n})$ and $B_j^n = \psi^{-1}(B_{y_j^n})$; here (as before) A_y and B_y are determined by the relation $g^{-1}(y) = A_y \times B_y$. For $(x,\lambda) \in \mathcal{X} \times \Lambda$, define

$$\bar{c}_n(x,\lambda) = \begin{cases} \nu_2(B_j^n)/\nu_1(B_j^n) & \text{if } \nu_1(B_j^n) > 0, \\ 1 & \text{if } \nu_1(B_j^n) = \nu_2(B_j^n) = 0, \end{cases} \tag{4.3.18}$$

$$\bar{c}(x,\lambda) = \limsup_{n\to\infty} \bar{c}_n(x,\lambda),$$

where j is determined by the relation $g^n(x,\lambda) = y_j^n$.

A direct computation verifies that, for any probability measure P on $\mathcal{X}$,

$$\bar{c}_n = E^{P\times\nu_1}\left[\frac{\nu_2(d\lambda)}{\nu_1(d\lambda)}\middle| \mathcal{F}_n \vee \mathcal{B}_n\right]. \tag{4.3.19}$$

Indeed, to show this it is sufficient to take any bounded measurable function, h, on $\mathcal{X} \times \mathcal{Y}$ and note that

$$\int_{\mathcal{X}} \int_{\Lambda} h(\varphi_n(x), g_n(x,\lambda)) \bar{c}_n(x,\lambda) P(dx) \nu_1(d\lambda)$$

$$= \sum_{j=1}^{\infty} \int_{A_j^n} h(\varphi_n(x), y_j^n) \frac{\nu_2(B_j^n)}{\nu_1(B_j^n)} P(dx) \nu_1(B_j^n)$$

$$= \int_{\mathcal{X}} \int_{\Lambda} h(\varphi_n(x), g_n(x,\lambda)) P(dx) \nu_2(d\lambda).$$

By (4.3.19) and Condition (i), $\bar{c}_n$ is a uniformly integrable martingale on $(\mathcal{X} \times \Lambda, (\mathcal{F}_n \vee \mathcal{G}_n)_{n \geq 1}, P \times \nu_1)$, for every P. Hence $\bar{c}_n$ converges to $\bar{c}$ with $P \times \nu_1$-measure 1 for every P, and satisfies

$$\bar{c}_n = E^{P \times \nu_1}[\bar{c} | \mathcal{F}_n \vee \mathcal{G}_n] \qquad \text{for every } n \geq 1. \tag{4.3.20}$$

Since we may take P to be concentrated on any single point $x \in \mathcal{X}$, we have actually shown that $\bar{c}_n(x,\lambda)$ converges to $\bar{c}(x,\lambda)$ for every $x \in \mathcal{X}$ and ν_1-almost every λ in Λ (where the exceptional set of ν_1-measure zero may depend on x).

It is obvious from the definition of $\bar{c}_n$ that $\bar{c}_n(x,\lambda)$ depends on x and λ only through $y_j^n = g_n(x,\lambda)$, and therefore that $\bar{c}_n$ is $\mathcal{G}_n$-measurable. It follows that $\bar{c}$ is measurable over $\bigvee_{n=m}^{\infty} \mathcal{G}_n$ for each m, and so (by Condition (ii)) $\bar{c}$ is measurable over $\mathcal{G}$. Since any $\mathcal{G}$-measurable function may be written as a Borel-measurable function of g, there exists some positive function, c, on $\mathcal{Y}$ with

$$\bar{c}(x,\lambda) = c \circ g(x,\lambda). \tag{4.3.21}$$

Now let h be bounded and measurable on $\mathcal{Y}$, let P be the probability measure on $\mathcal{X}$, and set

$$\bar{h}_n = E^{P \times \nu_2}[h \circ g | \mathcal{F}_n \vee \mathcal{G}_n]. \tag{4.3.22}$$

Again the martingale convergence theorem implies that $\bar{h}_n(x,\lambda)$ converges to $h \circ g(x,\lambda)$ for $P \times \nu_1$-almost every (x,λ), since $h \circ g$ is $\mathcal{G}$-measurable and Conditions (i) and (ii) imply that $\mathcal{G} \subset \bigvee_{n=1}^{\infty} \mathcal{G}_n \subset \bigvee_{n=1}^{\infty} (\mathcal{F}_n \vee \mathcal{G}_n)$. By Lebesgue's dominated convergence theorem, (4.3.20), and (4.3.19),

$$\begin{aligned}
\int_{\mathcal{Y}} hc \, dP^{g,\nu_1} &= \int_{\mathcal{X}} \int_{\Lambda} h\circ g \; c\circ g \; P(dx)\nu_1(d\lambda) \\
&= \lim_{n\to\infty} \int_{\mathcal{X}} \int_{\Lambda} \bar{h}_n \bar{c} \; P(dx)\nu_1(d\lambda) \quad \text{(by DCT)} \\
&= \lim_{n\to\infty} \int_{\mathcal{X}} \int_{\Lambda} \bar{h}_n \bar{c}_n \; P(dx)\nu_1(d\lambda) \quad \text{(by (4.3.20))} \\
&= \lim_{n\to\infty} \int_{\mathcal{X}} \int_{\Lambda} \bar{h}_n \; P(dx)\nu_2(d\lambda) \quad \text{(by (4.3.19))} \\
&= \int_{\mathcal{X}} \int_{\Lambda} h\circ g \; P(dx)\nu_2(d\lambda) \quad \text{(by DCT)} \\
&= \int_{\mathcal{Y}} h \, dP^{g,\nu_2}.
\end{aligned}$$

This verifies (4.3.17) and completes the proof. ||

Remark 1. In case it is possible to find $\{\varphi_n\}$ and $\{\psi_n\}$ so that $\mathcal{A}_n \subset \mathcal{A}_{n+1}$, Condition (i) in the theorem may be eliminated and Condition (ii) can be simplified to $\bigvee_{n=1}^{\infty} \mathcal{A}_n = \mathcal{A}$.

Remark 2. If φ_n and ψ_n are themselves countably-valued, then obviously g_n is also, so the theorem applies if Conditions (i) and (ii) are satisfied.

EXAMPLE 29 (continued). Letting <a> denote the closest integer to a (the larger integer in case of a tie), define

$$\varphi_n(x) = 2^{-n} <2^n x> \text{ and } \psi_n(\lambda) = 2^{-n} <2^n \lambda>.$$

It is straightforward to verify that Conditions (i) and (ii) in Theorem 7 are satisfied, and hence that Censoring Principle 2' follows in complete generality from the RLP (for this situation).

EXAMPLE 27 (continued). Let $\mathcal{P}_n = \{A_j^n\}_{j\le J_n}$ be a sequence of partitions of the unit sphere (in R^n) into finitely many Borel sets such that $\mathcal{P}_{n+1}$ refines $\mathcal{P}_n$ and

$$\lim_{n\to\infty} \max_{j\le J_n} \operatorname{diam}(A_j^n) = 0.$$

Let $\{\xi_j^n\}$ be a collection of points such that $\xi_j^n \in A_j^n$, and define

$$\varphi_n(x) = i2^{-n}\xi_j^n \quad \text{if} \quad i \leq 2^n|x| < i+1 \text{ and } x/|x| \in A_j^n,$$

$$\psi_n(\rho) = k2^{-n} \quad \text{if} \quad k \leq 2^n\rho < k+1.$$

Again the Conditions (i) and (ii) of Theorem 7 are easily verified, so that this censoring mechanism is also generally irrelevant.

4.3.4 Informative Censoring

It is, of course, not always the case that the censoring mechanism or distribution can be ignored. There are very few instances of fixed censoring wherein the mechanisms can be labeled informative, so we will concentrate in this section on random censoring.

The most common reason for being unable to ignore the censoring distribution, ν, in random censoring is dependence of the random variable X and the random censoring variable λ. In Example 29, for instance, one may have a non-cancer death which occurred because cancer substantially lowered overall health. Indeed in competing risk theory, in general, dependence between X and the censoring variables may be the rule rather than the exception. Such dependence makes Censoring Principle 2 inapplicable, and indeed $\ell_y(\theta)$ will typically depend upon ν in such situations. (The LP is still valid, of course.)

A second possible reason that the censoring distribution might be informative is that the censoring mechanism, g, might fail to be noninformative. As a very simple example, suppose the actual observation is

$$Y = g(X,\lambda) = X+\lambda,$$

where $X \in \mathcal{X} = (0,\infty)$ and $\lambda \in \Lambda = (0,\infty)$. It is easy to check that $g^{-1}(y)$ is *not* a product set in $\mathcal{X} \times \Lambda$ for any y, so that g clearly fails to be noninformative. For such g, $\ell_y(\theta)$ will typically depend on ν.

A third reason that ν might not be ignorable is that ν will often be unknown, and there could be some "prior" relationship between ν and θ. Again,

the notation of Section 3.5 is convenient here. Thus let θ stand for *all* unknown aspects of the situation and write $\theta = (\xi,\eta)$, where ξ is of interest and η is a nuisance variable (presumably containing unknown aspects of the distribution, ν, of λ). For instance, if $X \sim P_\xi$ and $\lambda \sim \nu_\eta$ are similar competing risks, there might well be suspected relationships between ξ and η which prevent ν_η from being ignored (even if X and λ are independent and g is noninformative). We will not repeat the discussion of Section 3.5 concerning when and why η (and hence ν_η) can be ignored in such situations.

A final kind of informative censoring should be mentioned, even though it is not censoring in the formal sense we have defined. This is censoring in which censored data is simply not observed or recorded. Thus, for the censoring mechanism described in (4.3.1) and (4.3.2), it could be the case that an $X_i > 10$ is not observed or even known to have existed. Such a situation is easily dealt with by recognizing that the relevant probability distribution of the *observed* X_i is the conditional distribution, given that $X_i \leq 10$. The censoring mechanism will usually enter into this conditional distribution in a nonignorable fashion, however.

Interestingly enough, this omission of data due to censoring can arise from the methods of *reporting* data (c.f. Dawid and Dickey (1977)). An obvious example is that of a trade journal which only publishes results of experiments which provide "significant" evidence according to some criteria. The data of interest, for a given issue, would be all data from experiments on that issue, but only that data leading to "significance" will become available; the rest will be censored. This is a very complicated problem, and it is not at all clear how to analyze the situation. The censoring of the journal clearly can not be ignored, however.

4.4 SIGNIFICANCE TESTING

4.4.1 Conflict with the LP

Significance testing of a hypothesis (used here in the sense of P-values, rather than α-level testing) is viewed by many as a crucial element

of statistics, yet it provides a startling and practically serious example of conflict with the LP. A significance test of the hypothesis H_0, that X has distribution P^0, proceeds by defining some statistic T(X), where large values of T supposedly cast doubt on H_0, and then calculating, for the given observation x, the significance level (or P-value) of x,

$$(4.4.1) \qquad p = P^0(T(X) \geq T(x)) = \int_{\{y: T(y) \geq T(x)\}} P^0(dy)$$

(i.e., the probability under P^0 of observing x or something more "extreme"). If this is small, then one supposedly doubts that H_0 could be true. General discussions of significance testing (including discussions of important practical issues such as "real" versus "statistical" significance) can be found in Edwards, Lindman, and Savage (1963), Hacking (1965), Morrison and Henkel (1970), Edwards (1972), Cox and Hinkley (1974), Dempster (1974a,b), Pratt (1976,1977), Cox (1977), Barnard (1980), Good (1981), Barnett (1982), Berger (1985), Hall and Selinger (1986), and Berger and Delampady (1987).

A very common setting for significance testing is the parametric framework of testing H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$. Then the null distribution, P^0, is simply P_{θ_0} in our usual notation (or $f_{\theta_0}(\cdot)$ if densities exist). In this parametric setting it is clear that reporting significance levels violates the LP, since significance levels involve averaging over sample points other than just the observed x (see (4.4.1)). The extremely serious practical problems that can result are discussed in Section 4.4.2.

Significance testing is also frequently used when only a single model P^0 is being contemplated. Testing of fit to a specified model is a common example. Since only one probability distribution is then involved, there is no likelihood function; it is hence often argued that the LP cannot apply to such a situation. Arguments to the contrary will be given in Section 4.4.3.

4.4.2 Averaging Over "More Extreme" Observations

The logic behind including all data "more extreme" than the given x, when calculating p, is not particularly convincing. Consider the following

artificial example, related to an example in Cox (1958).

EXAMPLE 30. Suppose, under P_0 and P_1, respectively, that X has the distributions given in the following table.

x	0	1	2	3	4
$P_0(x)$	.75	.14	.04	.037	.033
$P_1(x)$	.70	.25	.04	.005	.005

If $T(x) = x$ were used as the test statistic for a significance test of either P_0 or P_1 (i.e., if large x were considered "extreme"), and if $x = 2$ were observed, then the significance level against P_0 alone would be

$$p_0 = P_0(X \geq 2) = .11,$$

while the significance level against P_1 alone would be

$$p_1 = P_1(X \geq 2) = .05.$$

(We are not thinking here of testing P_0 versus P_1; the focus is on comparing significance tests of each separate hypothesis.) Thus $x = 2$ would provide "significant evidence against P_1 at the 5% level," but would not even provide "significant evidence against P_0 at the 10% level."

The concern here, of course, is that were P_0 and P_1 being considered simultaneously as possible models, likelihood reasoning would argue that they are equally supported by $x = 2$; their likelihood ratio is then equal to one. When considered in isolation therefore, it is definitely strange that $x = 2$ provides such different significance levels for P_0 and P_1.

Jeffreys (1961) clearly exposed the questionable logic behind significance levels, stating

> "...a hypothesis which may be true may be rejected because it has not predicted observable results which have not occurred."

In the example here, neither P_0 nor P_1 "predicts" that $x = 3$ or $x = 4$ will occur, and indeed they do not occur, but P_1 would be rejected at the 5% level, while P_0 would not, because P_1 "predicts" these *unobserved* results even less than P_0.

Questionable logic could perhaps be overlooked if it made little difference in practice, but here the averaging over other observations will virtually *always* have a profound effect. Consider the following example from Edwards, Lindman, and Savage (1963).

EXAMPLE 30.1. Suppose $X = (X_1,\ldots,X_n)$, where the X_i are i.i.d. $\mathcal{N}(\theta,\sigma^2)$, σ^2 known. The usual test statistic for testing H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$ is

$$T(X) = \sqrt{n}|\overline{X} - \theta_0|/\sigma,$$

where $\overline{X}$ is the sample mean. If $t = T(x)$ is the observed test statistic, the significance level is then

$$p = 2(1 - \Phi(t)),$$

where Φ is the standard normal c.d.f..

Consider, now, this testing scenario from a likelihood perspective. Were H_1 given by H_1: $\theta = \theta_1$, it would have been natural to use, as the comparative evidence for the two hypotheses, the observed likelihood ratio

$$L_{\theta_1} = f_{\theta_0}(x)/f_{\theta_1}(x).$$

Unfortunately, the actual H_1 consists of all $\theta \neq \theta_0$, making it difficult to define a true likelihood ratio, L, of H_0 to H_1. It seems clear, however, that a lower bound on L is

$$\underline{L} = f_{\theta_0}(x)/\sup_{\theta\neq\theta_0} f_\theta(x).$$

The evidence against H_0 is certainly no stronger than $\underline{L}$.

An easy calculation shows that, in this example,

$$\underline{L} = \exp\{-\tfrac{1}{2}t^2\}.$$

The following table gives values of $\underline{L}$ for various t, and also gives the significance levels associated with these t. (The $\underline{L}g$ row is discussed later.)

Table 1. Likelihood Ratio Bounds and Significance Levels

t	1.645	1.960	2.576	3.291
p	.10	.05	.01	.001
$\underline{L}$	.258	.146	.036	.0044
$\underline{L}g$	.644	.409	.123	.018

The surprise here is that $\underline{L}$ is much larger than p. When p is .05 for instance, $\underline{L}$ is .146, indicating that the data provides *no more* than 1 to 7 evidence against H_0.

$\underline{L}$ itself can be argued to be misleadingly small because it is based on maximizing the "likelihood of H_1." More reasonable is to use, as the "likelihood of H_1", an average of $f_\theta(x)$ over all $\theta \neq \theta_0$. This leads to a *weighted likelihood ratio*

$$Lg = f_{\theta_0}(x) / \int_{\{\theta \neq \theta_0\}} f_\theta(x)\, g(\theta) d\theta,$$

where g is some density (or "weight function"). A Bayesian would choose g to be the conditional prior density on H_1, in which case Lg would be the *Bayes factor*.

Regardless of interpretation, one can gain insight into the impact of such evidence measures by calculating lower bounds on Lg over reasonable classes of g. For instance, in Berger and Sellke (1987) it is shown that for *any* density g which is a nonincreasing function of $|\theta-\theta_0|$, Lg is at least as large as $\underline{L}g$, given in the last row of Table 1. The indication is thus that, when p = .05 say, the evidence against H_0 is actually no stronger than 1 to $2\frac{1}{2}$. (And if one tried "natural" functions g, one would find that Lg is typically 1 or more when p = .05; see, e.g., Jeffreys (1961).)

The above example is quite disturbing. It indicates that the classical statistician and the conditionalist will often reach very different conclusions with the same data, precisely because one averages over all "extreme" sample points while the other uses only the observed data. (Berger and Sellke (1987) specifically show that this averaging is the source of the

discrepancy.) Furthermore, the discrepancy between significance levels and conditional measures of evidence (e.g., $\underline{L}$, Lg or $\underline{Lg}$, the posterior probability of H_0, and even conditional frequentist measures -cf. Berger and Sellke (1987)) has been shown to hold in a huge variety of significance testing problems involving a "precise" hypothesis. (H_0 need not be a point null for the discrepancy to arise - see Berger and Sellke, 1987, and Berger and Delampady, 1987 - but if H_0 is, say, a one-sided hypothesis, then the discrepancy may not arise - see Casella and Berger, 1987.) Note also that this discrepancy is very related (but not identical) to "Jeffreys's Paradox" or "Lindley's Paradox". These issues are explored, in depth, in Edwards, Lindman, and Savage (1963), Berger and Sellke (1987) and Berger and Delampady (1987). Other relevant works include Lindley (1957, 1977), Jeffreys (1961), DeGroot (1973), Dempster (1974b), Dickey (1977), Smith and Spiegelhalter (1980), Good (1981, 1984), Shafer (1982), Zellner (1984), Berger (1985), Delampady and Berger (1987), and Delampady (1986a,b).

One defense of averaging over other observations (and at the same time an attack on the LP) that is sometimes advanced is the claim that it is necessary to consider what observations *might have* occurred. It is, however, a misconception to believe that the LP fails to do this. Indeed, in determining the likelihood function (or family of distributions for X), it is crucial to consider and compare the possible x that might be observed. Once this has been done, however, and the data obtained, the LP states that only the observed $\ell_x(\theta)$ is needed.

4.4.3 Testing A Single Null Model

When only P^0 has been formulated, it has been argued that significance testing does not violate the LP because nothing resembling a likelihood function exists. Although correct in a certain formal sense, there are several weaknesses to the argument.

Perhaps the most serious weakness follows from the observations in the previous section: if averaging over "extreme" sample points is

virtually *always* bad in testing a "precise" null when alternatives are given, it seems incredibly optimistic to believe that such averaging will be reasonable when alternatives are not given. The argument that "significance testing is the only available statistical procedure" is hardly persuasive when it is known that this available statistical procedure is bad for testing precise hypotheses.

A second weakness of the argument that only P^0 exists is that implicit alternatives to P^0 often are present. Indeed, alternatives must enter, at least informally, into the choice of the test statistic $T(x)$. For instance, in Example 30 it seems justifiable to use $T(x) = x$ to measure "extreme" only if the alternatives that one has in mind are, say, alternatives which are stochastically larger than P_0 (so that a large x tends to support the alternatives more than it tends to support P_0.) As another example of the implicit presence of alternatives, consider chi-square testing of fit.

EXAMPLE 30.2. Consider a statistical experiment in which n independent and identically distributed random quantities $X_1, X_2, \ldots, X_n$ are observed from a distribution F. It is desired to conduct a significance test of the hypothesis H_0: $F = F_0$, where F_0 is a specified distribution. A common test procedure, when no alternatives are specified, is the chi-square test of fit.

Chi-Square Test Procedure: First, a partition $\{a_i\}_{i=0}^m$ of the real line is selected. Then the sample frequencies of the n observations in the cells of the partition are calculated. Let $\underset{\sim}{z} = (z_1, \ldots, z_m)^t$ denote these frequencies; thus z_i = number of X_i's in $(a_{i-1}, a_i]$. Define

$$\theta_i = F(a_i) - F(a_{i-1}) = P^F(a_{i-1} < X \leq a_i),$$

$$\theta_i^0 = F_0(a_i) - F_0(a_{i-1}) = P^{F_0}(a_{i-1} < X \leq a_i),$$

and

$$\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_m)^t, \quad \underset{\sim}{\theta}^0 = (\theta_1^0, \ldots, \theta_m^0)^t.$$

Then the chi-square test procedure is to calculate the test statistic

$$t = \sum_{i=1}^{m} \frac{(z_i - n\theta_i^0)^2}{n\theta_i^0},$$

and approximate the significance level by

$$p = P(\chi^2_{m-1} \geq t),$$

where χ^2_{m-1} is a chi-square random variable with m-1 degrees of freedom.

The implied alternatives here arise from the fact that $\underset{\sim}{Z}$ has a *Multinomial* $(n,\underset{\sim}{\theta})$ distribution, so that basing the test on $\underset{\sim}{z}$ is equivalent to acknowledging the test to be that of H_0: $\underset{\sim}{\theta} = \underset{\sim}{\theta}^0$ versus H_1: $\underset{\sim}{\theta} \neq \underset{\sim}{\theta}^0$. (Use of t can be argued to further imply that the alternatives, $\underset{\sim}{\theta} \neq \underset{\sim}{\theta}^0$, are roughly ordered in plausibility according to $\eta = \sum_{i=1}^{m} (\theta_i - \theta_i^0)^2/\theta_i^0$, so that one is really testing H_0: $\eta = 0$ versus H_1: $\eta > 0$.) But this is a parametric problem with specified alternatives (and hence a likelihood function) so that LP-compatible testing methods can apply. Indeed, in Delampady and Berger (1987) it is shown that the same type of difficulty for significance testing, that was discussed in Section 4.4.2, exists here: the significance level is typically much smaller than sensible conditional measures of the evidence for H_0.

The above argument, that there are implicit alternatives in significance testing, can actually be given a quite general formal foundation. It has previously been mentioned that the actual sample space $\mathcal{X}$ will be discrete in practice. But then, as discussed in Section 3.6.1, even the set $\{P_\theta\}$ of *all* distributions on $\mathcal{X}$ actually results in a definable likelihood function. Furthermore, a significance test of P^0 can be identified with a test of H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$, where $P_{\theta_0} = P^0$. Thus the LP can apply, and argues against the use of significance levels.

Although formally correct, we do not ascribe much practical importance to this last argument, because the class of *all* alternatives to P^0 is typically much too big to suggest a sensible analysis. In practice, some consideration of the type of alternatives that are expected is necessary, even in classical significance testing. In choosing a test statistic T(x),

for instance, we earlier observed that it is often necessary to consider alternatives when defining "extreme." It has been argued that it may be easier to guess a reasonable T, reflecting intuitive judgements as to which observations support H_0 and which support alternatives, than to attempt explicit consideration of alternatives and construction of T by, say, likelihood ratio comparisons of P^0 with the alternatives. The argument that one can do better by use of intuition, than by explicit consideration of important relevant features of a problem (here, the alternatives), is difficult to refute, but is an argument that we would feel very uncomfortable having as a basis for our approach to science and understanding. Even more troubling is the fact that significance testing allows one to "hide" this use of personal intuition. Thus, while Pratt (1965) admits that consideration of alternatives can be hard and a source of controversy in many situations dealt with by significance testing, he argues that

> "Computing a P-value runs the danger of hiding this real uncertainty and legitimate disagreement behind a screen of irrelevant precision."

As a final point, it has been extensively argued (cf. Hacking (1965)) that one can never really reject P^0 until one has something better, namely another model P^1 which is both "reasonable" and better supported by the data. In Example 30, for instance, the observation $x = 2$ is quite unlikely to occur under P_0, but it is equally unlikely to occur under P_1; thus if P_0 and P_1 are *known* to be the only possibilities, then $x = 2$ provides no evidence against P_0. Thus consideration of alternatives is imperative if one actually seeks to reject P^0.

4.4.4 Conclusions

What is to be concluded about significance testing? First of all, it should be admitted that, as the significance level (or P-value) decreases, the evidence against H_0 will be increasing (assuming that T has been chosen

appropriately). Indeed, in a few special situations (primarily one-sided testing situations) the significance level can correspond to a reasonable conditional (Bayesian) measure of the validity of H_0 (cf. Jeffreys (1961), Pratt (1965), DeGroot (1973), Fraser and Mackay (1976), Dickey (1977), Zellner (1982), and Casella and Berger (1987)). In general though, the magnitude of a significance level need bear no relationship (from problem to problem) to the actual amount of evidence against H_0, and significance levels in testing precise hypotheses are typically so misleadingly small that their use for actually rejecting a hypothesis is strongly contraindicated.

Although a given significance level can mean vastly different things in different situations, it can be argued that, through frequent use in various situations, insight into its true strength of evidence against H_0 can be obtained. This is perhaps true: capable people can become very good at doing tasks with grossly inadequate tools. This is not to say, however, that better tools should be ignored or, more importantly, that inexperienced people will do well with the inadequate tools.

One possibly valid use of significance testing is to provide an alert that further investigation (in particular consideration of alternatives) is needed. As Barnard (1981) says

> "The question to be answered is whether the feature (T(x)) presented is so improbable on H_0 as to justify the effort involved in exercising our imagination to produce an hypothesis that could account for it."

There is no guarantee from a small significance level that P^0 is wrong (i.e., that an alternative hypothesis can be found which is substantially more supported by the data), but without a small significance level there may be no need to look past P^0. This use of significance testing can be argued to be important even to Bayesians, as extensively discussed in Box (1980): for a given model and prior, the marginal (or predictive) density of X can be

used to conduct a significance test which could alert one to question the model or prior.

Of course, even this use of significance testing as an alert could be questioned, because of the matter of averaging over unobserved x. It is hard to see what else could be done with P^0 *alone*, however, and it is sometimes argued that time constraints preclude consideration of alternatives. This may occasionally be true, but is probably fairly rare. Even cursory consideration of alternatives and a few rough likelihood ratio calculations will tend to give substantially more insight than will a significance level, and will usually not be much more difficult than sensibly choosing T and calculating the significance level. (See also Dempster (1974b).) Admittedly, such an approach will be somewhat imprecise, but what is the advantage of "irrelevant precision"?

4.5 RANDOMIZATION ANALYSIS

4.5.1 Introduction

In classical finite population sampling (or survey sampling) and randomization testing, the randomization in the experimental design (used to select the sample or allocate treatments) is a dominant factor in the construction of measures of evidence about θ. These measures are pre-experimental in nature, and their use directly violates the LP and RLP. (The outcome of the randomization is usually known, and hence averaging over samples or treatment allocations that might have occurred is supposedly irrelevant.) Hence, belief in the LP would have a profound effect on one's view of these areas of statistics.

Perhaps not surprisingly, it is in these areas, so drastically affected by the LP, that some of the strongest intuitive arguments against the LP can be raised. The issues involved are very complex, so much so that all we can hope to do is skim the surface of the subject. Indeed, we will essentially restrict ourselves to a defence of the LP in a few simple examples, trying to establish, as plausible, the argument that anything

sensible in randomization analysis is sensible precisely because it has a sensible interpretation from a likelihood viewpoint.

Although our main emphasis will not be on criticizing randomization analysis, it is important to keep several issues in mind. First, randomization analysis can clearly be very silly conditionally, if followed blindly. Even proper randomization can result (by bad luck) in treatment groups unbalanced with respect to unanticipated (but observed to be important) covariates, or in a sample which is clearly unrepresentative of the population, and yet classical randomization analysis does not treat such situations any differently than situations where the outcome of the randomization is "good". Thus, if one randomly samples from the population of voters in a survey on preference in the next Presidential election and finds that, unfortunately, all members of the random sample happen to be Republican, it is permissible (classically) to ignore this fact and proceed with the usual analysis. A second problem with randomization analysis (or at least randomization testing) is that it is often implemented through significance testing, and the serious concerns of the previous section then apply. The third, and most important, problem is that randomization analysis *does violate* the LP. In murky situations, where intuition stumbles, it seems especially necessary to depend on foundations.

Because of the above (and various specific) criticisms of randomization analysis, such analysis is usually advanced, not as an always sound way of proceeding, but as the most useful practical method of obtaining a reasonable answer. We will try to argue that the case for this is weak, at best.

Of course, even though we argue that the basis of randomization analysis is fundamentally in error, many of the specific procedures used in survey sampling and randomization testing are perfectly satisfactory. (If so, however, it is probably because they have some sensible interpretation consistent with the LP.) Also, the value of randomization itself, in treatment allocation and the choice of a sample, is not being addressed here.

Such randomization is often argued to be valuable (even by many conditionalists) in helping to reduce systematic effects that perhaps might unwittingly be introduced by the experimental design or sampling plan. An experiment in which randomization is used properly will, most of the time, turn out to be reasonably balanced with respect to experimentally induced (and unanticipated) covariates. Randomization also helps greatly in convincing others, who do not have access to the experimental setup or data, that no systematic biases were present. Employing measures of evidence based on the randomization probabilities is an entirely different matter, however. Indeed, a conditionalist will not only ignore the randomization probabilities, since the outcome of the randomization is known, but will also check to see that balance with respect to important new covariates was indeed obtained.

4.5.2 Finite Population Sampling

A typical classical setup is that of having a population $\mathcal{Y} = \{y_1,\ldots,y_N\}$ of N units, where each unit y_i can be represented as a vector $y_i = (u_i,v_i)$, u_i representing a label (or other known information) about the unit and v_i representing something unknown (but observable). It is desired to infer something about $\theta = (v_1,\ldots,v_N) \in \Theta$, from a sample $\mathcal{Y}_s = (y_{i_1}, y_{i_2},\ldots,y_{i_m})$, which is a subset of $\mathcal{Y}$; here $s = \{i_1,\ldots,i_m\} \subset I = \{1,\ldots,N\}$ indicates which units from the population are selected to be part of the sample. Note that it is typically also possible to use the known labels $\underset{\sim}{u} = (u_1,\ldots,u_N)$ in making inferences about θ. Let $\mathcal{S}$ denote the collection of all subsets of I, and suppose P is a probability distribution on $\mathcal{S}$. A procedure $\delta(\mathcal{Y}_s,\underset{\sim}{u})$ is to be used, and some criterion function $L(\delta(\mathcal{Y}_s,\underset{\sim}{u}),\theta)$ employed. Finally, the overall statistical procedure (P,δ), by which it is meant that s will be chosen according to the probability distribution P on $\mathcal{S}$ and $\delta(\mathcal{Y}_s,\underset{\sim}{u})$ will be used, is evaluated classically by the frequentist measure of performance

$$R(P,\delta,\theta) = \sum_{s\in\mathcal{S}} L(\delta(\mathcal{Y}_s,\underset{\sim}{u}),\theta)P(s).$$

EXAMPLE 31. Suppose it is desired to estimate the population total $\lambda = \sum_{i=1}^{N} v_i$, using an estimator $\delta(\mathcal{Y}_s, \underset{\sim}{u})$ and squared error loss $L(\delta(\mathcal{Y}_s, \underset{\sim}{u}), \theta) = (\delta(\mathcal{Y}_s, \underset{\sim}{u}) - \lambda)^2$. Suppose $P(s) = 1/\binom{N}{n}$ for all $s \in \mathcal{S}$ of size n, corresponding to selection of a simple random sample of size n. The estimator

$$\delta(\mathcal{Y}_s, \underset{\sim}{u}) = \frac{N}{n} \sum_{j=1}^{n} v_{i_j} \tag{4.5.1}$$

(recall that $\mathcal{Y}_s = ((u_{i_1}, v_{i_1}), \ldots, (u_{i_n}, v_{i_n}))$) is unbiased in the sense that

$$\sum_{s \in \mathcal{S}} \delta(\mathcal{Y}_s, \underset{\sim}{u}) P(s) = \lambda,$$

and hence $R(P, \delta, \theta)$ can be considered to be the variance of the procedure (P, δ), were it repeatedly used.

To investigate this situation from the viewpoint of likelihood, note that the only randomness here is in the generation of s, and hence that

$$P_\theta(\mathcal{Y}_s) = \begin{cases} P(s) & \text{if } \mathcal{Y}_s \in \Omega_\theta \\ 0 & \text{otherwise,} \end{cases} \tag{4.5.2}$$

where Ω_θ is the set of all possible vectors, $\mathcal{Y}_s$, which could arise as samples for the given $\underset{\sim}{u}$ and θ. (Note that the implicit sample space is the union, over all θ, of such Ω_θ.) Thus the likelihood function for θ, when $\mathcal{Y}_s$ is observed, is simply

$$\ell(\theta) = P(s) I_{\Lambda(\mathcal{Y}_s)}(\theta), \tag{4.5.3}$$

where (for $\mathcal{Y}_s = ((u_{i_1}, v_{i_1}), \ldots, (u_{i_m}, v_{i_m}))$)

$$\Lambda(\mathcal{Y}_s) = \{\theta \in \Theta: \text{ for } j = 1, \ldots, m, \text{ the } i_j \text{ component of } \theta \text{ equals } v_{i_j}\}.$$

Since $\ell(\theta)$ is constant for $\theta \in \Lambda(\mathcal{Y}_s)$, it conveys no information about θ, other than that the part of θ observed (in $\mathcal{Y}_s$) is known. This is deemed by some to be a failure of the LP, in that the statistical procedure is thought to provide considerable information about that part of θ not observed in $\mathcal{Y}_s$, call it θ^*.

The likelihood (or maybe Bayesian) view is indeed, that the data contains no inherent information about θ^*, and that the only way of infering anything about θ^* is to relate it somehow to the observed sample. Various relationships which might be deemed reasonable are:

(i) All v_i are thought to be similar, and the labels u_i contain no information. Of the many ways to model this, a simple (often too simple) possibility is to presume that the v_i are independent observations from a $\mathcal{N}(\mu,\tau^2)$ distribution. Then estimate μ and τ^2, using the sample $\mathcal{Y}_s$, and infer whatever is desired about θ. In the situation of Example 31, the answers would be essentially the same as the classical answers.

(ii) Suppose the v_i are thought to be linearly related to the u_i, say

$$v_i = \alpha + \beta u_i + \varepsilon_i,$$

where the ε_i are presumed to have some distribution. Clearly a quite different analysis would be appropriate.

(iii) Suppose two distinct similar groups within the population can be identified from $\mathcal{Y}_s$. Knowledge about each group can be obtained from $\mathcal{Y}_s$, as in (i), and the proportion of each group in the population estimated. (Of course, a stratified sample would probably have been desirable had the groups been identifiable solely from the labels.)

(iv) Suppose it is felt that the sample does not look typical of the remainder of the population. (An unlucky sample was drawn, or the sample revealed an unanticipated bias in the sampling plan.) It is not clear what to do, but it certainly cannot be right to proceed with a classical analysis, as if the sample was satisfactory.

In the situations above, classical sampling theorists would, of course, recommend different procedures for the various presumed models. The point of the discussion is to indicate that the data, $\mathcal{Y}_s$, really doesn't say anything about θ^*, unless there is some background information relating the

data to the population. It might be argued that, even when nothing is known about the population, a simple random sample will probably produce a representative subset of the population, so that an estimator such as (4.5.1) is reasonable for the population total. We do not disagree, but judge that (4.5.1) is then reasonable precisely because the sample is thought to be representative, in which case (4.5.1) would be justifiable from a variety of Bayesian arguments. The randomization *may* help to convince one that the sample is representative, but, once convinced of that fact, there is no further need to consider the sample selection probabilities.

Modeling the population is often called the superpopulation approach to survey sampling. Although we have presented it as Bayesian in nature, the modeling of the population can also be argued to be as "objective" as any modeling usually done in statistics (cf. the discussion by Royall of Basu (1971)), in which case one can argue that a directly meaningful likelihood function for the superpopulation parameters will exist. To a Bayesian, the choice of a model is just part of the prior specification (and often the most important and uncertain part), so the distinction seems unnecessary.

This discussion has assumed that the selection probabilities, P(s), are known. If they are partially unknown and depend on θ or on an informative nuisance parameter (see Section 3.5) they could be relevant to conclusions about θ. Rubin (1984) addresses this issue, distinguishing between "ignorable" and "nonignorable" sample selection mechanisms, and raises the related point that the P(s) may be useful as crude covariates in certain situations of stratified sampling.

Another issue that has been raised is the possibility of involving the P(s) by purposely ignoring the randomization outcome. Indeed, Rao (1971) argues that one can obtain an "informative" likelihood function by ignoring the labels u_i in the sample $\mathcal{Y}_s$. The available data is then only $\underset{\sim}{v}$, an m-vector of the observed v_i, with no record of which elements of the population it is associated with. It is easy to calculate, using (4.5.2) and

(4.5.3), that the likelihood function corresponding to $\underset{\sim}{y}$ is

$$(4.5.4) \qquad \ell(\theta) = \sum_{\text{all } \underset{\sim}{y}_s \text{ of size } m} P(s) I_{\Lambda(\underset{\sim}{y}_s)}(\theta).$$

This likelihood function may seem to contain more information about θ. In Example 31, for instance, it is easy to see that, if N/m is an integer, the M.L.E. for θ is any vector containing N/m copies of $\underset{\sim}{y}$. The M.L.E. for λ would thus be (4.5.1).

In discussing the reasonableness of the above proposal, it is important to first note that ignoring data is often a sensible practical necessity, as the following example indicates.

EXAMPLE 32. Suppose we observe (X,Y) having a joint density $f(x|\theta)g(y|\theta)$ (i.e., X and Y are independent), but that f is known while g is completely unknown. If we have very little prior information about g, so little that y conveys no clear knowledge about θ, then basing the analysis on x alone seems reasonable. Of course, ignoring y can be viewed as a formal violation of the LP, since it essentially involves integrating y out of the joint density of X and Y. It is not a violation of the spirit of the LP, however, providing $\ell_X(\theta) = f(x|\theta)$ is felt to be reasonably close to what would have been obtained were y included (say, by putting a prior distribution on g and integrating out over this prior). Further discussion and references on this issue can be found in Pratt (1965), who calls X an "insufficient statistic," and in Berger (1983).

While ignoring data may often be a practical necessity, there is a crucial difference between doing so in Example 32 and doing so in the sample survey problem. In Example 32 an *unknown* element g was eliminated by ignoring data, while Rao (1971) suggests replacing the *known* likelihood function in (4.5.3) by the version in (4.5.4) that would result if the labels in s were ignored. No real simplification is involved in the latter situation; indeed (4.5.4) seems more complicated than (4.5.3). In some

situations a non-Bayesian likelihood analysis of (4.5.4) may seem easier than a similar analysis of (4.5.3), but such is probably only the case in simple situations like that of Example 31 where P(s) is constant, (and then direct reasoning of a model construction or Bayesian nature with (4.5.3) is also easy). And it is easy to construct examples where the use of (4.5.4) with highly variable P(s) can give completely unreasonable answers for particular observed y_s.

We have barely touched the surface of survey sampling. Deeper discussions of these issues and other references can be found in Godambe (1966, 1982a, 1982b), Cornfield (1969), Basu (1969, 1971, 1978), Ericson (1969), Kalbfleisch and Sprott (1969), Rao (1971), Royall (1971, 1976), Godambe and Thompson (1976), Smith (1976), Cassel et. al. (1977), and Thompson (1980). A particularly convincing case for the Bayesian view can be found in Basu (1978).

4.5.3 Randomization Testing

Randomization testing was introduced by Fisher (cf. Fisher (1960)) and was further developed by Kempthorne and others. (See Kempthorne and Folkes (1971) and Basu (1980) for some of these developments and other references). The basis of randomization testing is using the randomization mechanism involved in treatment allocation to experimental units to form probability assessments of evidence. The following simple example exhibits the key features of the approach. See Basu (1980), and the discussants thereof, for a more general discussion.

EXAMPLE 33. In an experiment, n independent pairs of matched subjects $\{(s_1^0,s_1^1),\ldots,(s_n^0,s_n^1)\}$ are to be utilized to compare two treatments, T_0 (the "standard") and T_1 (the "new treatment"). Within each pair, the two treatments are randomly assigned: let r_i equal 0 or 1 as treatment T_0 or T_1, respectively, is assigned to s_i^0 (so that treatment $T_{(1-r_i)}$ is assigned to s_i^1), and define $\underset{\sim}{r} = (r_1,\ldots,r_n)$. Note that $P(r_i = 0) = P(r_i = 1) = \frac{1}{2}$. The result

of the experiment will be a vector $\underset{\sim}{X} = (X_1,\ldots,X_n)$, where, for the ith pair,

$$X_i = \begin{cases} 0 & \text{if } T_0 \text{ is judged to have worked better} \\ 1 & \text{if } T_1 \text{ is judged to have worked better.} \end{cases}$$

(For simplicity of discussion, we assume that equality of treatments is not a possible observation, and that only the crude measures X_i are observable.)

Randomization testing, here, would involve consideration of the hypothesis (H_0) that the treatments have an identical effect, in the sense that a given subject in each pair, say subject $S_i^{\delta_i}$ ($\delta_i = 0$ or 1), would do best no matter which treatment it received. It is easy to check that H_0 can be written mathematically as

$$(4.5.5) \qquad H_0: \; X_i = (r_i+\delta_i)_{\text{mod } 2} \quad \text{for } i = 1,\ldots,n.$$

Also, letting $\underset{\sim}{\delta} = (\delta_1,\ldots,\delta_n)$, it is clear that, *pre-experimentally*, $\underset{\sim}{X}$ has density (under H_0) $f_{\underset{\sim}{\delta}}(\underset{\sim}{x}) = 2^{-n}$ (since there is only one assignment $\underset{\sim}{r}$ which will match $\underset{\sim}{x}$ to $\underset{\sim}{\delta}$, and each $\underset{\sim}{r}$ has probability 2^{-n} of occurring).

Suppose that it is desired to perform a significance test of H_0 against the one-sided alternative that T_1 is a better treatment than T_0. The natural test statistic would be $X = \sum_{i=1}^{n} X_i$, with large values of X providing evidence against H_0. The significance level (or P-value) of an observation, $\underset{\sim}{x}$, would then be

$$\alpha = P_{H_0,\underset{\sim}{\delta}}(X \geq x = \sum_{i=1}^{n} x_i) = \sum_{j=x}^{n} \binom{n}{j} 2^{-n}.$$

If, for example, all $x_i = 1$, then $\alpha = 2^{-n}$ which, for large n, would seem to cast doubt on H_0.

The pre-experimental measure of evidence, α, in the above example is based on the randomization probabilities. Since the actual randomization outcome $\underset{\sim}{r}$ becomes known, however, conditional reasoning would argue that such probabilities are irrelevant. A conditional analysis of the problem might go as follows.

EXAMPLE 33 (continued). Because of the pairing (and the randomization) it might be deemed reasonable to pretend that the subjects within each pair are identical. If the pairs can be considered to be a random sample from the entire population of pairs, and θ denotes the (hypothetical) proportion of the population for which treatment T_1 would be better than T_0, then one could write the joint density of $\underset{\sim}{x}$ and $\underset{\sim}{r}$ as

$$f_\theta(\underset{\sim}{x},\underset{\sim}{r}) = 2^{-n}\theta^x(1-\theta)^{n-x}.$$

A likelihood analysis could then be performed, based on this (binomial) likelihood for θ. (Of course, a significance test of $\theta = \frac{1}{2}$ would give the same result as the randomization analysis, and we will argue that this is really why the randomization analysis is, at all, sensible.)

The randomization mechanism plays no direct role in the above likelihood argument. Indeed, the use of randomization is limited to making more believable the assumption that the paired subjects are equivalent: the randomization hopefully eliminates the possibility of experimenter induced bias that might be introduced by, say, giving treatment T_0 to the subjects (perhaps subconsciously) thought to be healthiest. It might be argued, by some, that the classical randomization analysis seems intuitively more sensible than the modeled likelihood analysis. The following illustration of biased randomization (as discussed in Basu (1980)) casts doubt on the validity of such an argument.

EXAMPLE 33 (continued). Suppose the treatments are assigned by a randomization mechanism having the property that the subjects S_i^0 (independently) receive treatment T_0 with probability $\frac{1}{4}$ and treatment T_1 with probability $\frac{3}{4}$. Suppose, further, that the randomization outcome happens to be that each S_i^0 receives treatment T_0, and the experimental outcome happens to be that each $x_i = 1$. If the null hypothesis is true, then it must be the case that $\delta_i = 1$ for all i (see (4.5.5)). But it follows that the significance level against H_0 is

$$\alpha = P_{H_0, \underset{\sim}{\delta} = (1,\ldots,1)}(X \geq x = n) = P(\text{all } r_i = 0) = 4^{-n}.$$

This significance level seems misleadingly low, due to the "unlikely" randomization outcome. The evidence against H_0 certainly seems no stronger than it would have been had an unbiased randomizer been used. The modeled likelihood analysis would, of course, be unaffected by the use of the biased randomizer. Thus it seems that the randomization analysis may be rather suspect, unless it corresponds to a sensible modeled likelihood analysis.

As with finite population sampling, the likelihood approach tends to involve further modeling of the situation under investigation. While to some extent unappealing (more assumptions must be introduced), there seems to be little choice. In Example 33, if one were not comfortable in treating the subjects within a pair as identical, or the pairs as representative of the population, then the randomization analysis would also be very suspect. (If it so happened that a certain subject in each pair could be identified as "healthier", a careful investigation of the matchups of treatments and subjects would be indicated.) Extensive discussions of these issues can be found (with other references) in Savage et. al. (1962), Hill (1970), Good (1976), Rubin (1978), Lindley and Novick (1981), and especially Basu (1980).