1.6. Foundations

We have defined a variety of expected losses, and decision principles based upon them, without discussing the advantages and disadvantages of each. Such discussion will actually be a recurring feature of the book, but in this section some of the most fundamental issues will be raised. The bulk of the section will be devoted to perhaps the most crucial issue in this discussion (and indeed in statistics), the conditional versus frequentist controversy, but first we will make a few comments concerning the common *misuse* of classical inference procedures to do decision problems. It should be noted that, while easy mathematically, many of the conceptual ideas in this foundations section are *very* difficult. This is a section that should frequently be reread as one proceeds through the book.

1.6.1. Misuse of Classical Inference Procedures

The bulk of statistics that is taught concerns classical inference procedures, and so it is only natural that many people will try to use them to do everything, even to solve clear decision problems. One problem with such use of inference procedures has already been mentioned, namely their failure to involve perhaps important prior and loss information. As another example (cf. Example 1) the loss in underestimation may differ substantially from the loss in overestimation, and any estimate should certaintly take this into account. Or, in hypothesis testing, it is often the case that the loss from an incorrect decision increases as a function of the "distance" of θ from the true hypothesis (cf. Example 1 (continued) in Subsection 4.4.3); this loss cannot be correctly measured by classical error probabilities.

One of the most commonly misused inference procedures is hypothesis testing (or significance testing) of a point null hypothesis. The following example indicates the problem.

EXAMPLE 8. A sample $X_1, \ldots, X_n$ is to be taken from a $\mathcal{N}(\theta, 1)$ distribution. It is desired to conduct a size $\alpha = 0.05$ test of H_0: $\theta = 0$ versus H_1: $\theta \neq 0$. The usual test is to reject H_0 if $\sqrt{n}|\bar{x}| > 1.96$, where $\bar{x}$ is the sample mean.

Now it is unlikely that the null hypothesis is ever exactly true. Suppose, for instance, that $\theta = 10^{-10}$, which while nonzero is probably a meaningless difference from zero in most practical contexts. If now a very large sample, say $n = 10^{24}$, is taken, then with extremely high probability $\bar{X}$ will be within 10^{-11} of the true mean $\theta = 10^{-10}$. (The standard deviation of $\bar{X}$ is only 10^{-12}.) But, for $\bar{x}$ in this region, it is clear that $10^{12}|\bar{x}| > 1.96$. Hence the classical test is virtually certain to reject H_0, even though the true mean is negligibly different from zero. This same phenomenon exists no matter what size $\alpha > 0$ is chosen and no matter how small the difference, $\varepsilon > 0$, is betweeen zero and the true mean. For a large enough sample size, the classical test will be virtually certain to reject.

The point of the above example is that it is meaningless to state only that a point null hypothesis is rejected by a size α test (or is rejected at significance level α). We *know* from the beginning that the point null hypothesis is almost certainly not exactly true, and that this will always be confirmed by a large enough sample. What we are really interested in determining is whether or not the null hypothesis is approximately true (see Subsection 4.3.3). In Example 8, for instance, we might really be interested in detecting a difference of at least 10^{-3} from zero, in which case a better null hypothesis would be $H_0: |\theta| \leq 10^{-3}$. (There are certain situations in which it is reasonable to formulate the problem as a test of a point null hypothesis, but even then serious questions arise concerning the "final precision" of the classical test. This issue will be discussed in Subsection 4.3.3.)

As another example of this basic problem, consider standard "tests of fit," in which it is desired to see if the data fits the assumed model. (A typical example is a test for normality.) Again it is virtually certain that the model is not exactly correct, so a large enough sample will almost always reject the model. The problem here is considerably harder to correct than in Example 8, because it is much harder to specify what an "approximately correct" model is.

A historically interesting example of this phenomenon (told to me by Herman Rubin) involves Kepler's laws of planetary motion. Of interest is his first law, which states that planetary orbits are ellipses. For the observational accuracy of Kepler's time, this model fit the data well. For todays data, however, (or even for the data just 100 years after Kepler) the null hypothesis that orbits are ellipses would be rejected by a statistical significance test, due to perturbations in the orbits caused by planetary interactions. The elliptical orbit model is, of course, essentially correct, the error caused by perturbations being minor. The concern here is that an essentially correct model can be rejected by too accurate data if statistical significance tests are blindly applied without regard to the actual size of the discrepancies.

The above discussion shows that a "statistically significant" difference between the true parameter (or true model) and the null hypothesis can be an unimportant difference practically. Likewise a difference that is not significant statistically can nevertheless be very important practically. Consider the following example.

EXAMPLE 9. The effectiveness of a drug is measured by $X \sim \mathcal{N}(\theta, 9)$. The null hypothesis is that $\theta \leq 0$. A sample of 9 observations results in $\bar{x} = 1$. This is not significant (for a one-tailed test) at, say, the $\alpha = 0.05$ significance level. It is significant at the $\alpha = 0.16$ significance level, however, which is moderately convincing. If 1 were a practically important difference from zero, we would certainly be very interested in the drug. Indeed if we had to make a decision solely on the basis of the given data, we would probably decide that the drug was effective.

The above problems are, of course, well recognized by classical statisticians (since, at least, Berkson (1938)) who, while using the framework of testing point null hypotheses, do concern themselves with the real import of the results. It seems somewhat nonsensical, however, to deliberately formulate a problem wrong, and then in an adhoc fashion explain the final results in more reasonable terms. Also, there are unfortunately many users of statistics who do not understand the pitfalls of the incorrect classical formulations.

One of the main benefits of decision theory is that it forces one to think about the correct formulation of a problem. A number of decision-theoretic alternatives to classical significance tests will be introduced as we proceed, although no systematic study of such alternatives will be undertaken.

1.6.2. The Frequentist Perspective

On the face of it, it may seem rather peculiar to use a risk (or any other frequentist measure such as confidence, error probabilities, bias, etc.) in the report from an experiment, since they involve averaging the performance of a procedure over all possible data, while it is known *which data* occurred. In this section we will briefly discuss the motivation for using frequentist measures.

Although one can undoubtedly find earlier traces, the first systematic development of frequentist ideas can be found in the early writings of J. Neyman and E. Pearson (cf. Neyman (1967)). The original driving force behind their frequentist development seemed to be the desire to produce measures which did not depend on θ, or any prior knowledge about θ. The method of doing this was to consider a procedure $\delta(x)$ and some criterion function $L(\theta, \delta(x))$ and then find a *number* $\bar{R}$ such that repeated use of δ would yield average long run performance of at least $\bar{R}$.

EXAMPLE 10. For dealing with standard univariate normal theory problems, consider the usual 95% confidence rule for the unknown mean θ,

$$\delta(x) = (\bar{x} - ts, \bar{x} + ts),$$

where $\bar{x}$ and s are the sample mean and standard deviation, respectively, and t is the appropriate percentile from the relevant t distribution. Suppose that we measure the performance of δ by

$$L(\theta, \delta(x)) = 1 - I_{\delta(x)}(\theta) = \begin{cases} 0 & \text{if } \theta \in \delta(x), \\ 1 & \text{if } \theta \notin \delta(x). \end{cases}$$

Note that, if this is treated as a decision-theoretic loss, the risk becomes (including the unknown standard deviation σ as part of the parameter)

$$R((\theta, \sigma), \delta) = E^X_{\theta,\sigma} L(\theta, \delta(X)) = P_{\theta,\sigma}(\delta(X) \text{ does not contain } \theta) = 0.05.$$

The idea now is to imagine that we will use δ repeatedly on a series of (independent, say) normal problems with means θ_i, standard deviations σ_i, and data $X^{(i)}$. It is then an easy calculation, using the law of large numbers, to show that (with probability one)

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} L(\theta_i, \delta(X^{(i)})) = 0.05 \equiv \bar{R}, \tag{1.1}$$

no matter what sequence of (θ_i, σ_i) is encountered.

The above frequentist motivation carries considerable appeal. As statisticians we can "proclaim" that the universal measure of performance that is to be reported with δ is $\bar{R} = 0.05$, since, on the average in repeated use, δ will fail to contain the true mean only 5% of the time. This appealing motivation for the frequentist perspective was formalized as the Confidence Principle by Birnbaum (see Cox and Hinkley (1974) and Birnbaum (1977) for precise formulations). Other relevant works of Neyman on this issue are Neyman (1957 and 1977).

Two important points about the above frequentist justification should be stressed. These are: (i) the motivation is based on repeated use of δ for *different* problems; and (ii) a bound $\bar{R}$ on performance must be found which applies to *any* sequence of parameters from these different problems. The elimination of either of these features considerably weakens the case for the frequentist measure. And, indeed, the risk $R(\theta, \delta)$, that we have so far considered, seems to violate *both* of these conditions; it is defined as the repeated average loss if one were to use δ on a series of data from the *same* problem (since θ is considered fixed), and a report of the function $R(\theta, \delta)$ has not eliminated dependence on θ.

Several justifications for $R(\theta, \delta)$ can still be given in terms of the "primary motivation," however. The first is that risk dominance of δ_1 over δ_2 will usually imply that δ_1 is better than δ_2 in terms of the primary motivation. The second is that $R(\theta, \delta)$ may have an upper bound $\bar{R}$, and, if so, this can typically be shown to provide the needed report for the primary motivation. To see the problem in using just $R(\theta, \delta)$, consider the following example.

EXAMPLE 11. Consider testing the simple null hypothesis H_0: $\theta = \theta_0$ versus the simple alternative hypothesis H_1: $\theta = \theta_1$. If the loss is chosen to be "0–1" loss (see Subsection 2.4.2), the risk function of a test δ turns out to be given by $R(\theta_0, \delta) = \alpha_0 = P_{\theta_0}$ (Type I error) and $R(\theta_1, \delta) = \alpha_1 = P_{\theta_1}$ (Type II error). Suppose now that one always uses the most powerful test of level $\alpha_0 = 0.01$. This would allow one to make the frequentist statement, upon rejecting H_0, "my procedure ensures that only 1% of true null hypotheses will be rejected."

Unfortunately, this says *nothing* about how often one errs when rejecting. For instance, suppose $\alpha_1 = 0.99$ (admittedly terrible Type II error probability, but useful for making the point) and that the null and alternative

parameter values occur equally often in repetitive use of the test. (Again, we are imagining repeated use of the $\alpha_0 = 0.01$, $\alpha_1 = 0.99$ most powerful test on a sequence of different simple versus simple testing problems.) Then it can easily be shown that *half* of all rejections of the null will actually be in error. And *this* is the "error" that really measures long run performance of the test (when rejecting). Thus one cannot make useful statements about the actual error rate incurred in repetitive use, without a satisfactory bound on $R(\theta, \delta)$ for all θ.

Other justifications for $R(\theta, \delta)$ can be given involving experimental design and even "Bayesian robustness" (see Subsection 1.6.5 and Section 4.7). It will be important, however, to bear in mind that all these justifications are somewhat secondary in nature, and that assigning inherent meaning to $R(\theta, \delta)$, as an experimental report, is questionable. For more extensive discussion of this issue, see Berger (1984b), which also provides other references.

1.6.3. The Conditional Perspective

The conditional approach to statistics is concerned with reporting data-specific measures of accuracy. The overall performance of a procedure δ is deemed to be of (at most) secondary interest; what is considered to be of primary importance is the performance of $\delta(x)$ for the *actual data* x that is observed in a given experiment. The following simple examples show that there can be a considerable difference between conditional and frequentist measures.

EXAMPLE 12. Suppose that X_1 and X_2 are independent with identical distribution given by

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

where $-\infty < \theta < \infty$ is unknown. The procedure (letting $X = (X_1, X_2)$)

$$\delta(X) = \begin{cases} \text{the point } \frac{1}{2}(X_1 + X_2) & \text{if } X_1 \neq X_2, \\ \text{the point } X_1 - 1 & \text{if } X_1 = X_2, \end{cases}$$

is easily seen to be a frequentist 75% confidence procedure of smallest size (i.e., $P_\theta(\delta(X) = \theta) = 0.75$ for all θ). However, a conditionalist would reason as follows, depending on the particular x observed: if x has $x_1 \neq x_2$, then we *know* that $\frac{1}{2}(x_1 + x_2) = \theta$ (since one of the observations *must be* $\theta - 1$ and the other *must be* $\theta + 1$), while, if $x_1 = x_2$, the data fails to distinguish in any way between the two possible θ values $x_1 - 1$ and $x_1 + 1$. Hence, conditionally, $\delta(x)$ would be 100% certain to contain θ if $x_1 \neq x_2$, while if $x_1 = x_2$ it would be 50% certain to contain θ.

Careful consideration of this example will make the difference between the conditional and frequentist viewpoints clear. The overall performance of δ, in any type of repeated use, would indeed be 75%, but this arises because half the time the *actual* performance will be 100% and half the time the *actual* performance will be 50%. And, for any given application, one *knows* whether one is in the 100% or 50% case. It clearly would make little sense to conduct an experiment, use $\delta(x)$, and actually report 75% as the measure of accuracy, yet the frequentist viewpoint suggests doing so. Here is another standard example.

EXAMPLE 13. Suppose X is 1, 2, or 3 and θ is 0 or 1, with X having the following probability density in each case:

	x		
	1	2	3
$f(x\|0)$	0.005	0.005	0.99
$f(x\|1)$	0.0051	0.9849	0.01

The classical most powerful test of H_0: $\theta = 0$ versus H_1: $\theta = 1$, at level $\alpha = 0.01$, concludes H_1 when $X = 1$ or 2, and this test also has a Type II error probability of 0.01. Hence, a standard frequentist, upon observing $x = 1$, would report that the decision is H_1 and that the test had error probabilities of 0.01. This certainly *gives the impression* that one can place a great deal of confidence in the conclusion, but is this the case? Conditional reasoning shows that the answer is sometimes no! When $x = 1$ is observed, the likelihood ratio between $\theta = 0$ and $\theta = 1$ is $(0.005)/(0.0051)$ which is very close to one. To a conditionalist (and to most other statisticians also), a likelihood ratio near one means that the data does very little to distinguish between $\theta = 0$ and $\theta = 1$. Hence the conditional "confidence" in the decision to conclude H_1, when $x = 1$ is observed, would be only about 50%. (Of course $x = 1$ is unlikely to occur, but, when it does, should not a sensible answer be given?)

The next example is included for historical reasons, and also because it turns out to be a key example for development of the important Likelihood Principle in the next subsection. This example is a variant of the famous Cox (1958) conditioning example.

EXAMPLE 14. Suppose a substance to be analyzed can be sent either to a laboratory in New York or a laboratory in California. The two labs seem equally good, so a fair coin is flipped to choose between them, with "heads" denoting that the lab in New York will be chosen. The coin is flipped and

comes up tails, so the California lab is used. After a while, the experimental results come back and a conclusion and report must be developed. Should this conclusion take into account the fact that the coin could have been heads, and hence that the experiment in New York might have been performed instead? Common sense (and the conditional viewpoint) cries no, that only the experiment *actually performed* is relevant, but frequentist reasoning would call for averaging over all possible data, even the possible New York data.

The above examples were kept simple to illustrate the ideas. Many complex and common statistical situations in which conditioning seems very important can be found in Berger and Wolpert (1984) and the references therein. An example is the use of observed, rather than expected, Fisher information (see Subsection 4.7.8). Examples that will be encountered in this book include hypothesis testing (see Subsection 4.3.3), several decision-theoretic examples, and the very important example of optional stopping, which will be considered in Section 7.7. (The conditional viewpoint leads to the conclusion that many types of optional stopping of an experiment can be ignored, a conclusion that can have a drastic effect on, for instance, the running of clinical trials.)

Savage (1962) used the term *initial precision* to describe frequentist measures, and used the term *final precision* to describe conditional measures. Initially, i.e., before seeing the data, one can only measure how well δ is likely to perform through a frequentist measure, but after seeing the data one can give a more precise final measure of performance. (The necessity for using at least partly frequentist measures in designing experiments is apparent.)

The examples above make abundantly clear the necessity for consideration of conditioning in statistics. The next question, therefore, is—What kind of conditional analysis should be performed? There are a wide variety of candidates, among them Bayesian analysis, fiducial analysis (begun by R. A. Fisher, see Fisher (1935)), various "likelihood methods" (cf. Edwards (1972) and Hinde and Aitkin (1986)), structural inference (begun by D. A. S. Fraser, see Fraser (1968)), pivotal inference (see Barnard (1980)), and even a number of conditional frequentist approaches (see Kiefer (1977a) or Berger (1984b, 1984c)). Discussion of these and other conditional approaches (as well as related conditional ideas such as that of a "relevant subset") can be found in Barnett (1982), Berger and Wolpert (1984), and Berger (1984d), along with many references. In this book we will almost exclusively use the Bayesian approach to conditioning, but a few words should be said about the conditional frequentist approaches because they can provide important avenues for generalizing the book's frequentist decision theory to allow for conditioning (cf. Kiefer (1976, 1977a) and Brown (1978)).

Kiefer (1977a) discussed two types of conditional frequentist approaches, calling them "conditional confidence" and "estimated confidence." The

idea behind conditional confidence is to use frequentist measures, but conditioned on subsets of the sample space. Thus, in Example 12, it would be possible to condition on $\{x: x_1 = x_2\}$ and $\{x: x_1 \neq x_2\}$, and then use frequentist reasoning to arrive at the "correct" measures of confidence. And in Example 14, one could condition on the outcome of the coin flip.

The estimated confidence approach does not formally involve conditioning, but instead allows the reported confidence to be data dependent. Thus, in Example 12, one could report a confidence of 100% or 50% as $x_1 \neq x_2$ or $x_1 = x_2$, respectively. The frequentist aspect of this estimated confidence approach is that the average *reported* performance in repeated use will be equal to the *actual* average performance, thus satisfying the primary frequentist motivation. For a rigorous statement of this, and a discussion of the interesting potential that estimated confidence has for the frequentist viewpoint, see Kiefer (1977a) and Berger (1984b and 1984c).

1.6.4. The Likelihood Principle

In attempting to settle the controversies surrounding the choice of a paradigm or methodology for statistical analysis, many statisticians turn to foundational arguments. These arguments generally involve the proposal of axioms or principles that any statistical paradigm should follow, together with a logical deduction from these axioms of a particular paradigm or more general principle that should be followed. The most common such foundational arguments are those that develop axioms of "rational behavior" and prove that any analysis which is "rational" must correspond to some form of Bayesian analysis. (We will have a fair amount to say about these arguments later in the book.) A much simpler, and yet profoundly important, foundational development is that leading to the Likelihood Principle. Indeed the Likelihood Principle, by itself, can go a long way in settling the dispute as to which statistical paradigm is correct. It also says a great deal about how one should condition.

The Likelihood Principle makes explicit the natural conditional idea that *only* the actual observed x should be relevant to conclusions or evidence about θ. The key concept in the Likelihood Principle is that of the likelihood function.

Definition 11. For observed data, x, the function $l(\theta) = f(x|\theta)$, considered as a function of θ, is called the *likelihood function.*

The intuitive reason for the name "likelihood function" is that a θ for which $f(x|\theta)$ is large is more "likely" to be the true θ than a θ for which $f(x|\theta)$ is small, in that x would be a more plausible occurrence if $f(x|\theta)$ were large.

The Likelihood Principle. *In making inferences or decisions about θ after x is observed, all relevant experimental information is contained in the likelihood function for the observed x. Furthermore, two likelihood functions contain the same information about θ if they are proportional to each other (as functions of θ).*

EXAMPLE 15 (Lindley and Phillips (1976)). We are given a coin and are interested in the probability, θ, of having it come up heads when flipped. It is desired to test H_0: $\theta = \frac{1}{2}$ versus H_1: $\theta > \frac{1}{2}$. An experiment is conducted by flipping the coin (independently) in a series of trials, the result of which is the observation of 9 heads and 3 tails.

This is not yet enough information to specify $f(x|\theta)$, since the "series of trials" was not explained. Two possibilities are: (1) the experiment consisted of a predetermined 12 flips, so that $X = [\#$ heads$]$ would be $\mathcal{B}(12, \theta)$; or (2) the experiment consisted of flipping the coin until 3 tails were observed, so that X would be $\mathcal{NB}(3, \theta)$. The likelihood functions in cases (1) and (2), respectively, would be

$$l_1(\theta) = f_1(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = (220)\theta^9(1-\theta)^3$$

and

$$l_2(\theta) = f_2(x|\theta) = \binom{n+x-1}{x}\theta^x(1-\theta)^n = (55)\theta^9(1-\theta)^3.$$

The Likelihood Principle says that, in either case, $l_i(\theta)$ is all we need to know from the experiment, and, furthermore, that l_1 and l_2 would contain the *same* information about θ since they are proportional as functions of θ. Thus we did not really need to know anything about the "series of trials"; knowing that independent flips gave 9 heads and 3 tails would, by itself, tell us that the likelihood function would be proportional to $\theta^9(1-\theta)^3$.

Classical analyses, in contrast, are quite dependent on knowing $f(x|\theta)$, and not just for the observed x. Consider classical significance testing, for instance. For the binomial model, the significance level of $x = 9$ (against $\theta = \frac{1}{2}$) would be

$$\begin{aligned}\alpha_1 = P_{1/2}(X \geq 9) &= f_1(9|\tfrac{1}{2}) + f_1(10|\tfrac{1}{2}) + f_1(11|\tfrac{1}{2}) + f_1(12|\tfrac{1}{2}) \\ &= 0.075.\end{aligned}$$

For the negative binomial model, the significance level would be

$$\begin{aligned}\alpha_2 = P_{1/2}(X \geq 9) &= f_2(9|\tfrac{1}{2}) + f_2(10|\tfrac{1}{2}) + \cdots \\ &= 0.0325.\end{aligned}$$

If significance at the 5% level was desired, the two models would thus lead to quite different conclusions, in contradiction to the Likelihood Principle.

Several important points, illustrated in the above example, should be emphasized. First the correspondence of information from proportional likelihood functions applies *only* when the two likelihood functions are for the *same* parameter. (In the example, θ is the probability of heads for the given coin on a single flip, and is thus defined independently of which experiment is performed. If l_1 had applied to one coin, and l_2 to a different coin, the Likelihood Principle would have had nothing to say.)

A second point is that the Likelihood Principle does *not* say that all information about θ is contained in $l(\theta)$, just that all *experimental* information is. There may well be other information relevant to the statistical analysis, such as prior information or considerations of loss.

The example also reemphasizes the difference between a conditional perspective and a frequentist type of perspective. The significance level calculations involve not just the observed $x = 9$, but also the "more extreme" $x \geq 10$. Again it seems somewhat peculiar to involve, in the evaluation, observations that have not occurred. No one has phrased this better than Jeffreys (1961):

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

Thus, in Example 15, the null hypothesis that $\theta = \frac{1}{2}$ certainly would not predict that X would be larger than 9, and indeed such values *do not occur.* Yet the probabilities of these unpredicted and not occurring observations are included in the classical evidence against the hypothesis.

Here is another interesting example (from Berger and Wolpert (1984)).

EXAMPLE 16. Let $\mathscr{X} = \{1, 2, 3\}$ and $\Theta = \{0, 1\}$, and consider experiments E_1 and E_2 which consist of observing X_1 and X_2, respectively, both having sample space $\mathscr{X}$ and the *same* unknown θ. The probability densities of X_1 and X_2 are (for $\theta = 0$ and $\theta = 1$)

	x_1		
	1	2	3
$f_1(x_1\|0)$	0.90	0.05	0.05
$f_1(x_1\|1)$	0.09	0.055	0.855

	x_2		
	1	2	3
$f_2(x_2\|0)$	0.26	0.73	0.01
$f_2(x_2\|1)$	0.026	0.803	0.171

If, now, $x_1 = 1$ is observed, the Likelihood Principle states that the information about θ should depend on the experiment only through $(f_1(1|0), f_1(1|1)) = (0.90, 0.09)$. Furthermore, since this is proportional to $(0.26, 0.026) = (f_2(1|0), f_2(1|1))$, the Likelihood Principle states that $x_2 = 1$ would provide the same information about θ as $x_1 = 1$. Another way of stating the Likelihood Principle for testing simple hypotheses, as here, is that the experimental evidence about θ is contained in the likelihood ratio

for the observed x. Note that the likelihood ratios for the two experiments are also the same when 2 is observed, and also when 3 is observed. Hence, no matter which experiment is performed, the *same* conclusion about θ should be reached for the given observation.

This example clearly indicates the startling nature of the Likelihood Principle. Experiments E_1 and E_2 are very different from a frequentist perspective. For instance, the test which accepts $\theta = 0$ when the observation is 1 and decides $\theta = 1$ otherwise is a most powerful test with error probabilities (of Type I and Type II, respectively) 0.10 and 0.09 for E_1, and 0.74 and 0.026 for E_2. Thus the classical frequentist would report drastically different information from the two experiments.

The above example emphasizes the important distinction between initial precision and final precision. Experiment E_1 is much more *likely* to provide useful information about θ, as evidenced by the overall better error probabilities (which are measures of initial precision). Once x is at hand, however, this initial precision is no longer relevant, and the Likelihood Principle states that whether x came from E_1 or E_2 is irrelevant. This example also provides a good testing ground for the various conditional methodologies that were mentioned in Subsection 1.6.3. For instance, either of the conditional frequentist approaches has a very hard time in dealing with the example.

So far we have not given any reasons why one *should* believe in the Likelihood Principle. Examples 15 and 16 are suggestive, but could perhaps be viewed as refutations of the Likelihood Principle by die-hard classicists. Before giving the axiomatic justification that exists for the Likelihood Principle, we indulge in one more example in which it would be very hard to argue against the Likelihood Principle.

EXAMPLE 17 (Pratt (1962)). "An engineer draws a random sample of electron tubes and measures the plate voltages under certain conditions with a very accurate voltmeter, accurate enough so that measurement error is negligible compared with the variability of the tubes. A statistician examines the measurements, which look normally distributed and vary from 75 to 99 volts with a mean of 87 and a standard deviation of 4. He makes the ordinary normal analysis, giving a confidence interval for the true mean. Later he visits the engineer's laboratory, and notices that the voltmeter used reads only as far as 100, so the population appears to be 'censored'. This necessitates a new analysis, if the statistician is orthodox. However, the engineer says he has another meter, equally accurate and reading to 1000 volts, which he would have used if any voltage had been over 100. This is a relief to the orthodox statistician, because it means the population was effectively uncensored after all. But the next day the engineer telephones and says, 'I just discovered my high-range voltmeter was not working the day I did the experiment you analyzed for me.' The statistician ascertains that the engineer

would not have held up the experiment until the meter was fixed, and informs him that a new analysis will be required. The engineer is astounded. He says, 'But the experiment turned out just the same as if the high-range meter had been working. I obtained the precise voltages of my sample anyway, so I learned exactly what I would have learned if the high-range meter had been available. Next you'll be asking about my oscilloscope.'"

In this example, two different sample spaces are being discussed. If the high-range voltmeter had been working, the sample space would have effectively been that of a usual normal distribution. Since the high-range voltmeter was broken, however, the sample space was truncated at 100, and the probability distribution of the observations would have a point mass at 100. Classical analyses (such as the obtaining of confidence intervals) would be considerably affected by this difference. The Likelihood Principle, on the other hand, states that this difference should have no effect on the analysis, since values of x which did not occur (here $x \geq 100$) have no bearing on inferences or decisions concerning the true mean. (A formal verification is left for the exercises.)

Rationales for at least some forms of the Likelihood Principle exist in early works of R. A. Fisher (cf. Fisher (1959)) and especially of G. A. Barnard (cf. Barnard (1949)). By far the most persuasive argument for the Likelihood Principle, however, was given in Birnbaum (1962). (It should be mentioned that none of these three pioneers were unequivocal supporters of the Likelihood Principle. See Basu (1975) and Berger and Wolpert (1984) for reasons, and also a more extensive historical discussion and other references. Also, the history of the concept of "likelihood" is reviewed in Edwards (1974).)

The argument of Birnbaum for the Likelihood Principle was a proof of its equivalence with two other almost *universally* accepted natural principles. The first of these natural principles is the sufficiency principle (see Section 1.7) which, for one reason or another, almost everyone accepts. The second natural principle is the (weak) conditionality principle, which is nothing but a formalization of Example 14. (Basu (1975) explicitly named the "weak" version.)

The Weak Conditionality Principle. *Suppose one can perform either of two experiments E_1 or E_2, both pertaining to θ, and that the actual experiment conducted is the* mixed *experiment of first choosing $J = 1$ or 2 with probability $\frac{1}{2}$ each (independent of θ), and then performing experiment E_J. Then the actual information about θ obtained from the overall mixed experiment should depend only on the experiment E_j that is actually performed.*

For a proof that sufficiency together with weak conditionality imply the Likelihood Principle in the case of discrete $\mathscr{X}$, see Birnbaum (1962) or Berger and Wolpert (1984); the latter work also gives a similar development

and proof in an extremely general probabilistic setting. The argument poses a serious challenge to all who are unwilling to believe the Likelihood Principle; the only alternatives are to reject the sufficiency principle (which would itself cause havoc in classical statistics) or to reject the weak conditionality principle—yet what could be more obvious?

There have been a number of criticisms of Birnbaum's axiomatic development, including concerns about the existence of the likelihood function(i.e., of $f(x|\theta)$), and even of the existence of "information from an experiment about θ." Also, some of the consequences of the Likelihood Principle are so startling (such as the fact that the Likelihood Principle implies that optional stopping of an experiment should usually be irrelevant to conclusions, see Section 7.7) that many statisticians simply refuse to consider the issue. Basu (1975), Berger and Wolpert (1984), and Berger (1984d) present (and answer) essentially all of the criticisms that have been raised, and also extensively discuss the important consequences of the Likelihood Principle (and the intuitive plausibility of these consequences).

It should be pointed out that the Likelihood Principle does have several inherent limitations. One has already been mentioned, namely that, in designing an experiment, it is obviously crucial to take into account all x that can occur; frequentist measures (though perhaps Bayesian frequentist measures) must then be considered. The situation in sequential analysis is similar, in that, at a given stage, one must decide whether or not to take another observation. This is essentially a design-type problem and, in making such a decision, it may be necessary to know more than the likelihood function for θ from the data observed up until that time. (See Section 7.7 for further discussion.) A third related problem is that of prediction of future observables, in which one wants to predict a future value of X. Again, there may be information in the data beyond that in the likelihood function for θ. Actually, the Likelihood Principle will apply in all these situations if θ is understood to consist of *all* unknowns relevant to the problem, including further random X, and not consist just of unknown model parameters. See Berger and Wolpert (1984) for discussion.

The final, yet most glaring, limitation of the Likelihood Principle is that it does not indicate how the likelihood function is to be used in making decisions or inferences about θ. One proposal has been to simply report the entire likelihood function, and to educate people in its interpretation. This is perhaps reasonable, but is by no means the complete solution. First of all, it is frequently also necessary to consider the prior information and loss, and the interaction of these quantities with the likelihood function. Secondly, it is not at all clear that the likelihood function, by itself, has any particular meaning. It is natural to attempt to interpret the likelihood function as some kind of probability density for θ. The ambiguity arises in the need to then specify the "measure" with respect to which it is a density. There are often many plausible choices for this measure, and the choice can have a considerable effect on the conclusion reached. This problem is

basically that of choosing a "noninformative" prior distribution, and will be discussed in Chapter 3.

Of the methods that have been proposed for using the likelihood function to draw conclusions about θ (see Berger and Wolpert (1984) for references), only the Bayesian approach seems generally appropriate. This will be indicated in the next section, and in Chapter 4. (More extensive such arguments can be found in Basu (1975) and Berger and Wolpert (1984).) It will also be argued, however, that a good Bayesian analysis may sometimes require a slight violation of the Likelihood Principle, in attempting to protect against the uncertainties in the specification of the prior distribution. The conclusion that will be reached is that analysis compatible with the Likelihood Principle is an ideal towards which we should strive, but an ideal which is not always completely attainable.

In the remainder of the book, the Likelihood Principle will rarely be used to actually do anything (although conditional Bayes implementation of it will be extensively considered). The purpose in having such a lengthy discussion of the principle was to encourage the "post-experimental" way of thinking. Classical statistics teaches one to think in terms of "pre-experimental" measures of initial precision. The Likelihood Principle states that this is an error; that one should reason only in terms of the actual sample and likelihood function obtained. Approaching a statistical analysis with this viewpoint in mind is a radical departure from traditional statistical reasoning. And note that, while the Likelihood Principle is the "stick" urging adoption of the conditional approach, there is also the "carrot" that the conditional approach often yields great simplification in the statistical analysis: it is usually much easier to work with just the observed likelihood function, rather than having to involve $f(x|\theta)$ for *all* x, as a frequentist must (see also Sections 4.1 and 7.7).

1.6.5. Choosing a Paradigm or Decision Principle

So far we have discussed two broad paradigms, the conditional and the frequentist, and, within each, a number of possible principles or methodologies that could be followed. As these various paradigms and decision principles are discussed throughout the book, considerable effort will be spent in indicating when the methods seem to work and, more importantly, when they do not. The impression that may emerge from the presentation is that statistics is a collection of useful methodologies, and that one should "keep an open mind as to which method to use in a given application." This is indeed the most common attitude among statisticians.

While we endorse this attitude in a certain practical sense (to be made clearer shortly), we do *not* endorse it fundamentally. The basic issue is—How can we *know* that we have a sensible statistical analysis? For example, how can we be certain that a particular frequentist analysis has not run

afoul of a conditioning problem? It is important to determine what *fundamentally* constitutes a sound statistical analysis, so that we then have a method of judging the practical soundness and usefulness of the various methodologies.

We have argued that this desired fundamental analysis must be compatible with the Likelihood Principle. Furthermore, we will argue in Chapter 4 that it is conditional Bayesian analysis that is the only fundamentally correct conditional analysis. From a practical viewpoint, however, things are not so clearcut, since the Bayesian approach requires specification of a prior distribution π, for θ, and this can never be done with complete assurance (see Section 4.7). Hence we will modify our position (in Section 4.7) and argue that the fundamentally correct paradigm is the "robust Bayesian" paradigm, which takes into account uncertainty in the prior.

Unfortunately, robust Bayesian analysis turns out to be quite difficult; indeed for many problems it is technically almost impossible. We thus run into the need for what Good (1983) calls "Type II rationality": when time and other realistic constraints in performing a statistical analysis are taken into account, the optimal analysis may be an analysis which is not rigorously justifiable (from, say, the robust Bayesian viewpoint). The employment of any alternative methodology should, however, be justified from this perspective, the justification being that one is in this way most likely to be "close to" the philosophically correct analysis.

With the above reasoning, we will be able to justify a number of uses of frequentist measures such as $R(\theta, \delta)$. Also, recall that partially frequentist reasoning is unavoidable in many statistical domains, such as design of experiments and sequential analysis. A final justification for consideration of $R(\theta, \delta)$ is that, whether we like it or not, the bulk of statistical analyses that will be performed will use prepackaged procedures. Although the primary concern should be to see that such procedures are developed so as to be conditionally sound, the fact that they will see repeated use suggests that verification of acceptable long-run performance would only be prudent. In spite of all these reasons, we would strongly argue that conditional (Bayesian) reasoning should be the primary weapon in a statistician's arsenal.

It should be noted that we did not attempt to justify use of frequentist measures on certain "traditional" grounds such as the desire for "objectivity" or avoidance of use of subjective inputs (such as prior information). Objectivity is clearly very difficult in decision theory, since one cannot avoid subjective choice of a loss function. Even more to the point, strong arguments can be made that one can *never* do truly objective (sensible) statistical analyses; analyses that have the appearance of objectivity virtually always contain hidden, and often quite extreme, subjective assumptions. (For instance, the choice of a model is usually a very sharp subjective input.) Some indications of this will be seen throughout the book, although for more thorough discussions (of this and the other foundational issues), see

Jeffreys (1961), Zellner (1971), Box and Tiao (1973), Good (1983), Jaynes (1983), Berger (1984a), and Berger and Wolpert (1984) (all of which also have other references).

With the exception of Chapters 3 and 4, the book will tend to emphasize methodologies based on $R(\theta, \delta)$. The reason is mainly historical: the bulk of existing statistical decision-theoretic methodology is frequentist in nature. We will often pause, however, to view things from the conditional perspective.

1.7. Sufficient Statistics

The concept of a sufficient statistic (due to Fisher (1920, 1922)) is of great importance in simplifying statistical problems. Intuitively, a sufficient statistic is a function of the data which summarizes all the available sample information concerning θ. For example, if an independent sample $X_1, \ldots, X_n$ for a $\mathcal{N}(\mu, \sigma^2)$ distribution is to be taken, it is well known that $T = (\bar{X}, S^2)$ is a sufficient statistic for $\theta = (\mu, \sigma^2)$. (Here $S^2 = \sum (X_i - \bar{X})^2/(n-1)$.)

It is assumed that the reader is familiar with the concept of sufficiency and with the methods of finding sufficient statistics. We will content ourselves here with a rather brief discussion of sufficiency, including a presentation of the major decision-theoretic result concerning sufficiency. (For an in-depth examination of sufficiency, see Huzurbazar (1976).) The following formal definition of sufficiency uses the concept of a conditional distribution, with which the reader is also assumed familiar.

Definition 12. Let X be a random variable whose distribution depends on the unknown parameter θ, but is otherwise known. A function T of X is said to be a *sufficient statistic* for θ if the conditional distribution of X, given $T(X) = t$, is independent of θ (with probability one).

For understanding the nature of a sufficient statistic and for the development of the decision-theoretic result concerning sufficiency, the concept of a partition of the sample space must be introduced.

Definition 13. If $T(X)$ is a statistic with range $\mathcal{J}$ (i.e., $\mathcal{J} = \{T(x): x \in \mathcal{X}\}$), the *partition of* $\mathcal{X}$ induced by T is the collection of all sets of the form

$$\mathcal{X}_t = \{x \in \mathcal{X}: T(x) = t\}$$

for $t \in \mathcal{J}$.

Note that if $t_1 \neq t_2$, then $\mathcal{X}_{t_1} \cap \mathcal{X}_{t_2} = \emptyset$, and also observe that $\bigcup_{t \in \mathcal{J}} \mathcal{X}_t = \mathcal{X}$. Thus $\mathcal{X}$ is divided up (or partitioned) into the disjoint sets $\mathcal{X}_t$.