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 certainly 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.