2

Optimal Procedures

2.1 Defining Optimal

As we saw in the preceding chapter, the professional statistician is responsible for choosing both the test statistic and the testing procedure. An amateur might hope to look up the answers in a book, or, as is all too commonly done, use the same statistical procedure as was used the time before, regardless of whether it continues to be applicable. But the professional is responsible for choosing the best procedure, the optimal statistic. The statistic we selected in the preceding chapter for testing the effectiveness of vitamin E seemed an obvious, intuitive choice. But is it the best choice? And can we prove it is? Intuition can so often be deceptive.

In this chapter, we examine the criteria that define an optimal testing procedure and explore the interrelationships among them.

2.1.1 Trustworthy

The most obvious desirable property of a statistical procedure is that it be trustworthy. If we are advised to make a particular decision, then we should be correct in doing so. Alas, our observations are stochastic in nature, so there may be more than one explanation for any given set of observations. The result is we never can rely 100% on the decisions we make. At best, they can be like politicians, trustworthy up to a point. We ask only that they confine themselves to small bribes and rake-offs, that they not bankrupt or betray the country.

In the example of the missing labels in the preceding chapter, we introduced a statistical test based on the random assignment of labels to treatments. Knowing in advance that the experiment could have any of $\binom{6}{3} = 20$ possible outcomes, we will reject the null hypothesis only if the obtained value of the test statistic is the maximum possible that could arise from only one permutation of the results. The test we derive is valid under very broad

assumptions. The data could have been drawn from a normal distribution or they could have come from some quite different distribution. To be valid at a given percent level, all that is required of our permutation test is that (under the hypothesis) the population from which the data in the treatment group are drawn be the same as that from which the untreated sample is taken.

This freedom from reliance on numerous assumptions is a big plus. The fewer the assumptions, the fewer the limitations, and the broader the potential applications of a test. But before statisticians introduce a test into their practice, they need to know a few more things about it, namely:

- Is it *exact*? That is, can we make an exact determination of the probability that we might make an error in rejecting a true hypothesis?
- How *powerful* a test is it? That is, how likely is it to pick up actual differences between treated and untreated populations? Is this test as powerful or more powerful than the test we are using currently?
- Is the test *admissible*? That is, is there no other test that is superior to it under all circumstances?
- How *robust* is the new test? That is, how sensitive is it to violations in the underlying assumptions and the conditions of the experiment?

2.1.2 Two Types of Error

It's fairly easy to reason from cause to effect—that is, if you have a powerful enough computer. Get the right formula (Boyle's Law, say), plug in enough values to enough decimal places, and out pops the answer. The difficulty with reasoning in the opposite direction, from effect to cause, is that more than one set of causes can be responsible for precisely the same set of effects. We can never be completely sure which set of causes is responsible. Consider the relationship between sex (cause) and height (effect). Boys are taller than girls. True? So that makes this new $6'2''$ person in our lives ... a starter on the women's volleyball team.

In real life, in real populations, there are vast differences from person to person. Some women are tall and some women are short. In Lake Wobegone, Minnesota, all the men are good looking and all the children are brighter than average. But in most other places in the world there is a wide range of talent and abilities. As a further example of this variation, consider that half an aspirin will usually take care of one person's headache while other people take two or three aspirin at a time and get only minimal relief.

Figure 2.1 below depicts the results of an experiment in which two groups were each given a "pain-killer." The first group got buffered aspirin; the second group received a new experimental drug. Each of the participants then provided a subjective rating of the effects of the drug. The ratings ranged from "got worse," to "much improved," depicted below on a scale of 0 to 4. Take a close look at Figure 2.1. Does the new drug represent an improvement over aspirin?