11

Coping with Disaster

In this chapter you receive practical guidelines for coping with the many catastrophes that confront the applied statistician:

- subjects who miss an appointment;
- subjects who disappear completely and mysteriously in the middle of an experiment;
- incomplete questionnaires;
- covariates after the fact;
- outlying observations whose extreme and questionable values suggest they may have been recorded incorrectly;
- off-scale and other censored values that cannot be determined with precision.

11.1 Missing Data

The effects of missing data depend upon the nature of the study and the type of analysis. In some instances, for example, in the analysis of the k-sample comparison by permutation means, missing data may have no effect upon the analysis other than to reduce the power of the test. In other, more complex designs, missing data may result in an unbalanced design in which several factors are confounded with one another. In most, though not all, of these latter cases, no special statistical procedures are required, *providing* we are careful in how we interpret the results. We must identify which effects are confounded with one another, a main effect with an interaction, say. In other studies, and one such example was examined in Section 7.7.1, we may have to abandon permutation and parametric procedures altogether and consider using the bootstrap.

The majority of experimental designs belong to the correctable category. We proceed with a permutation rather than a parametric analysis using a revised set of marginal constraints that reflect the actual rather than the

hoped-for sample sizes. And in analyzing the results, we acknowledge that one or more higher-order interactions may have contaminated the observed effects.

Consider an example we studied in Chapter 7, the effect of sunlight and fertilizer on crop yield. Suppose that one of the observations in the low sunlight, medium fertilizer group, the 22 noted in parentheses in the table below, is missing from the study.

Effect of sunlight and fertilizer on crop yield

	Fertilizer		
Sunlight	LO	MED	HI
	5	15	21
LO	10	(22)	29
	8	18	25
	6	25	55
HI	9	32	60
	12	40	48

The test statistic for the main effect of sunlight is the sum of the observations at the low level, $S = 23 + (15 + 18) + 75 = 131$. Such an extremely low value is found in only a small handful of the rearrangements in which we swap observations at random between the low and high groups. The number of rearrangements after correcting for the missing data item is $\binom{17}{8}$. The reduction from the hoped for $\binom{18}{8}$ rearrangements reduces the power of the test. But the reduction is irrelevant in this instance as we are rejecting the hypothesis. (Had we accepted the null hypothesis, we would have been forced to consider whether a larger sample size might have enabled us to detect an effect.)

A missing data item in only one of the groups means that the main effect of sunlight is partially confounded with the interaction between sunlight and fertilizer. But our common sense strengthened by a glance at the table tells us that the confounding also is irrelevant in this instance.

One other word of caution: A variety of software is available today to help you determine optimal sample size. But such software does not take into account the possibility of missing data, of failures in recruitment, and long-term retention. Always use the numbers such software provides as starting points, not as final estimates.

The preceding discussion was based on the implicit assumption that dropouts occur at random. If the dropout rate is directly related to the treatment, we must either abandon the study or modify our scoring system explicitly to account for the dropouts (see, for example, Entsuah [1990]).

A further example of using the permutation distribution to cope with missing data is given in Section 12.2.6. Section 12.5 details the use of the bootstrap when all other methods fail.