CHAPTER 1

VERY LARGE ARRAYS

1.1. APPLICATIONS

Very large arrays of data, that is, data sets for which the number of observations per subject may be an order of magnitude greater than the number of subjects that are observed, arise in genetics research (microarrays), neurophysiology (EEGs), and image analysis (ultrasound, MRI, fMRI, MEG, and PET maps, telemetry). Microarrays of as many as 22,000 genes may be collected from as few as 50 subjects. While EEG readings are collected from a relatively small number of leads, they are collected over a period of time, so that the number of observations per subject is equal to the number of leads times the number of points in time at which readings are taken. fMRI images of the brain can be literally four dimensional when the individual time series are taken into account.

Analyzing the Large Numbers of Variables in Biomedical and Satellite Imagery, First Edition.
Phillip I. Good.

In this chapter, we consider the problems that arise when we attempt to analyze such data, potential solutions to these problems, and our plan of attack in the balance of this book.

1.2. PROBLEMS

1. The limited number of subjects means that the precision of any individual observation is equally limited. If n is the sample size, the precision of any individual observation is roughly proportional to the square root of n.
2. The large number of variables means that it is almost certain that changes in one or several of them will appear to be statistically significant purely by chance.
3. The large number of variables means that missing and/or erroneously recorded data is inevitable.
4. The various readings are not independent and identically distributed; rather, they are interdependent both in space and in time.
5. Measurements are seldom Gaussian (normally distributed), nor likely to adhere to any other well-tabulated distribution.

1.3. SOLUTIONS

Solutions to these problems require all of the following.

Distribution-free methods—permutation tests, bootstrap, and decision trees—are introduced in Chapters 2, 6, and 7, respectively. Their application to very large arrays is the subject of Chapters 3, 6, and 8.

One might ask, why not use parametric tests? To which Karniski et al. [1994] would respond:

> Utilizing currently available parametric statistical tests, there are essentially four methods that are frequently used to attempt

to answer the question. One may combine data from multiple variables to reduce the number of variables, such as in principal component analysis. One may use multiple tests of single variables and then adjust the critical value.

One may use univariate tests, and then adjust the results for violation of the assumption of sphericity (in repeated measures design). Or one may use multivariate tests, *so long as the number of subjects far exceeds the number of variables*.

Methods for reducing the number of variables under review are also considered in Chapters 3, 5, and 8.

Methods for controlling significance levels and/or false detection rates are discussed in Chapter 5.

Chapter 4, on gathering and preparing data, provides the biomedical background essential to those who will be analyzing very large data sets derived from medical images and microarrays.