A MISSING INFORMATION PRINCIPLE: THEORY AND APPLICATIONS

TERENCE ORCHARD
and
MAX A. WOODBURY
DUKE UNIVERSITY MEDICAL CENTER

1. Introduction

The problem that a relatively simple analysis is changed into a complex one just because some of the information is missing, is one which faces most practicing statisticians at some point in their career. Obviously the best way to treat missing information problems is not to have them. Unfortunately circumstances arise in which information is missing and nothing can be done to replace it for one reason or another. In analogy to other accidents—we don't plan on accidents, nevertheless they do occur and safety measures must be aimed at palliating consequences as well as at prevention. Consequently, a great volume of literature has been produced, dealing with a number of specific situations. An indication of the content of many of these papers is given in the Appendix. In this paper we propose to try to present a general philosophy for dealing with the problem of missing information, and to give a method which will lead quite easily to maximum likelihood estimates of the parameters obtained from the incomplete data using as nearly as possible the same techniques as if the data were all present.

Our first simple use of the missing information principle resulted from a conversation in 1946 between Max A. Woodbury and C. W. Cotterman resulting from the latter's interest (Cotterman [20]) in estimating gene frequencies from phenotypic frequency data. The observation was made that if one has the genotypic *frequencies* NAA, NAB, NBA, NBB, NAO, NBO, NOB, and NOO of red blood cell genotypes, indicated by the second and third letters of each symbol, then the gene frequencies are easily computed. If N is the total of the above frequencies then the estimates would be

$$\begin{aligned} \hat{p}_A &= \frac{1}{2N}(2NAA + NAB + NBA + NAO + NOA), \\ \hat{p}_B &= \frac{1}{2N}(2NBB + NBA + NAB + NBO + NOB), \qquad (1.1) \\ \hat{p}_o &= \frac{1}{2N}(2NOO + NOA + NAO + NOB + NBO). \end{aligned}$$

Research conducted under PHS-NIH Grant GM 16725-02 from the National Institute of General Medical Sciences.