THE MODEL BASED (PREDICTION) APPROACH TO FINITE POPULATION SAMPLING THEORY

Richard M. Royall, Department of Biostatistics,
The Johns Hopkins University

Introduction

Estimating a finite population mean from a sample is equivalent to predicting the mean of the non-sample values. This view, that finite population inference problems are actually prediction problems, leads naturally to a theory in which prediction models, not sample selection probabilities, are central. This paper is an informal survey of that theory.

The first section describes the model-based approach and attempts to make clear how and why it differs from the prevailing (randomization-based) theory. This section is built around a simple example, which is used to illustrate various facets of the approach. The second section addresses the question "What has the model-based approach accomplished?" This is not an attempt to catalog significant contributions to model-based sampling theory, but to describe and interpret the general kinds of developments that have occurred. Finally, the third section consists of some brief observations on current research.

What Is Model-Based Sampling Theory?

Model-based sampling theory begins by recognizing that problems of estimating finite population characteristics are naturally expressed as prediction problems (Kalbfleisch and Sprott, 1969; Geisser, 1986, p. 163). For example, Figure 1 shows the data for a sample of $n = 32$ hospitals. For each sample hospital we know the number of beds (x) and we have observed the number of patients discharged (y) during a given month. If we must estimate how many patients were discharged from another hospital, say one with x = 400 beds, we might fit the dotted line in Figure 1. The slope of that line, the ratio of total sample discharges to total sample beds, shows that in sample hospitals there were 3.1 patients discharged per bed. Thus we might estimate that there were about 3.1 × 400 = 1240 patients discharged from the other hospital. More generally, to estimate how many patients were discharged from a set r of non-sample hospitals having a total of $\Sigma_r x_i$ beds, we might use 3.1 $\Sigma_r x_i$. Then to estimate the patient total for the entire population composed of the thirty-two hospitals in the sample s as well as those in r, we would simply add the observed total for the thirty-two sample hospitals, $\Sigma_s y_i$ to our estimate for those not observed, 3.1 $\Sigma_r x_i$.

Clearly this estimate of the population total is reasonable only if it is reasonable to assume that the hospitals in r are "like" the ones in s: if the sample hospitals are in the eastern United States while the r-hospitals are in France, then this estimate is certainly questionable. How can we formalize this reasoning,