EXHIBIT F

Sampling Techniques

third edition

WILLIAM G. COCHRAN

Professor of Statistics, Emeritus
Harvard University

JOHN WILEY & SONS
New York • Chichester • Brisbane • Toronto • Singapore

to Bet

Library of Congress Cataloging in Publication Data:

Cochran, William Gemmell, 1909–
Sampling techniques.

(Wiley series in probability and mathematical statistics)
Includes bibliographical references and index.
1. Sampling (Statistics) I. Title.
QA276.6.C6 1977 001.4'222 77–728
ISBN 0–471–16240–X

Printed in the United States of America

30 29 28

Printed and bound by Quinn - Woodbine, Inc.

For example, if a probability sample of the records of batteries in routine use in a large factory shows an average life $\hat{\mu} = 394$ days, with a standard error $\sigma_{\hat{\mu}} = 4.6$ days, the chances are 99 in 100 that the average life in the population of batteries lies between

$$\hat{\mu} = 394 - (2.58)(4.6) = 382 \text{ days}$$

and

$$\hat{\mu}_U = 394 + (2.58)(4.6) = 406 \text{ days}$$

The limits, 382 days and 406 days, are called lower and upper *confidence limits*. With a single estimate from a single survey, the statement "μ lies between 382 and 406 days" is not certain to be correct. The "99% confidence" figure implies that if the same sampling plan were used many times in a population, a confidence statement being made from each sample, about 99% of these statements would be correct and 1% wrong. When sampling is being introduced into an operation in which complete censuses have previously been used, a demonstration of this property is sometimes made by drawing repeated samples of the type proposed from a population for which complete records exist, so that μ is known (see, e.g., Trueblood and Cyert, 1957). The practical verification that approximately the stated proportion of statements is correct does much to educate and reassure administrators about the nature of sampling. Similarly, when a single sample is taken from each of a series of different populations, about 95% of the 95% confidence statements are correct.

The preceding discussion assumes that $\sigma_{\hat{\mu}}$, as computed from the sample, is known exactly. Actually, $\sigma_{\hat{\mu}}$, like $\hat{\mu}$, is subject to a sampling error. With a normally distributed variable, tables of Student's t distribution are used instead of the normal tables to calculate confidence limits for μ when the sample is small. Replacement of the normal table by the t table makes almost no difference if the number of degrees of freedom in $\sigma_{\hat{\mu}}$ exceeds 50. With certain types of stratified sampling and with the method of replicated sampling (section 11.19) the degrees of freedom are small and the t table is needed.

1.8 BIAS AND ITS EFFECTS

In sample survey theory it is necessary to consider biased estimators for two reasons.

1. In some of the most common problems, particularly in the estimation of ratios, estimators that are otherwise convenient and suitable are found to be biased.

2. Even with estimators that are unbiased in probability sampling, errors of measurement and nonreponse may produce biases in the numbers that we are able to compute from the data. This happens, for instance, if the persons who refuse to be interviewed are almost all opposed to some expenditure of public funds, whereas those who are interviewed are split evenly for and against.

Since the probability that the ith unit is drawn is $1/N$ at each draw, the variate t_i distributed as a binomial number of successes out of n trials with $p = 1/N$. Hence

$$E(t_i) = \frac{n}{N}, \qquad V(t_i) = n\left(\frac{1}{N}\right)\left(1 - \frac{1}{N}\right) \tag{2.33}$$

Jointly, the variates t_i follow a multinomial distribution. For this,

$$\text{Cov}\,(t_i t_j) = -\frac{n}{N^2} \tag{2.34}$$

Using (2.32), (2.33), and (2.34), we have, for sampling with replacement,

$$V(\bar{y}) = \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 \frac{n(N-1)}{N^2} - 2\sum_{i<j}^{N} y_i y_j \frac{n}{N^2}\right] \tag{2.35}$$

$$= \frac{1}{nN}\sum_{i=1}^{N}(y_i - \bar{Y})^2 = \frac{\sigma^2}{n} = \frac{N-1}{N}\frac{S^2}{n} \tag{2.36}$$

Consequently, $V(\bar{y})$ in sampling without replacement is only $(N-n)/(N-1)$ times its value in sampling with replacement. If instead of $\bar{y}$ the mean $\bar{y}_d$ of the different or distinct units in the sample is used as an estimate when sampling is with replacement, Murthy (1967) has shown that the leading term in the average variance of $\bar{y}_d$ is $(1 - f/2)S^2/n$, following work by Basu (1958) and Des Raj and Khamis (1958). In some applications the cost of measuring the distinct units in the sample may be predominating, so that the cost of the sample is proportional to the number of distinct units. In this situation, Seth and J. N. K. Rao (1964) showed that for given average cost, $V(\bar{y})$ in sampling without replacement is less than $V(\bar{y}_d)$ in sampling with replacement. They also prove the more general result that if $\bar{y}_d' = f(\nu)\bar{y}_d / Ef(\nu)$, where ν is the number of distinct units in the sample and $f(\nu)$ is a function of ν, then $V(\bar{y}) < V(\bar{y}_d')$ if $S^2 < N\bar{Y}^2$, a condition satisfied by nearly all populations encountered in sample surveys.

2.11 ESTIMATION OF A RATIO

Frequently the quantity that is to be estimated from a simple random sample is the ratio of two variables both of which vary from unit to unit. In a household survey examples are the average number of suits of clothes per adult male, the average expenditure on cosmetics per adult female, and the average number of hours per week spent watching television per child aged 10 to 15. In order to estimate the first of these items, we would record for the ith household ($i = 1, 2, \ldots, n$) the number of adult males x_i who live there and the total number of

This allocation is sometimes called *Neyman allocation*, after Neyman (1934), whose proof gave the result prominence. An earlier proof by Tschuprow (1923) was later discovered.

A formula for the minimum variance with fixed n is obtained by substituting the value of n_h in (5.26) into the general formula for $V(\bar{y}_{st})$. The result is

$$V_{min}(\bar{y}_{st}) = \frac{\left(\sum W_h S_h\right)^2}{n} - \frac{\sum W_h S_h^2}{N} \tag{5.27}$$

The second term on the right represents the fpc.

5.6 RELATIVE PRECISION OF STRATIFIED RANDOM AND SIMPLE RANDOM SAMPLING

If intelligently used, stratification nearly always results in a smaller variance for the estimated mean or total than is given by a comparable simple random sample. It is not true, however, that *any* stratified random sample gives a smaller variance than a simple random sample. If the values of the n_h are far from optimum, stratified sampling may have a higher variance. In fact, even stratification with optimum allocation for fixed total sample size may give a higher variance, although this result is an academic curiosity rather than something likely to happen in practice.

In this section a comparison is made between simple random sampling and stratified random sampling with proportional and optimum allocation. This comparison shows how the gain due to stratification is achieved.

The variances of the estimated *means* are denoted by V_{ran}, V_{prop}, and V_{opt}, respectively.

Theorem 5.8. If terms in $1/N_h$ are ignored relative to unity,

$$V_{opt} \le V_{prop} \le V_{ran} \tag{5.28}$$

where the optimum allocation is for fixed n, that is, with $n_h \propto N_h S_h$.

Proof.

$$V_{ran} = (1-f)\frac{S^2}{n} \tag{5.29}$$

$$V_{prop} = \frac{(1-f)}{n}\sum W_h S_h^2 = \frac{\sum W_h S_h^2}{n} - \frac{\sum W_h S_h^2}{N} \tag{5.30}$$

[from equation (5.8), section 5.3]

$$V_{opt} = \frac{\left(\sum W_h S_h\right)^2}{n} - \frac{\sum W_h S_h^2}{N} \tag{5.31}$$

[from equation (5.27), section 5.5]