The Annals of Statistics
1977, Vol. 5, No. 2, 419-428

BOOK REVIEWS

H. S. KONIJN, *Statistical Theory of Sample Survey Design and Analysis*. North Holland, Amsterdam, 1973, xv+429 pages, fl 93.60.

Review by J. SEDRANSK
State University of New York at Buffalo

1. Introduction. In the development of a theory of sampling from finite populations it is commonly assumed that there is a finite population of N distinguishable units labelled as $\mathbf{N} = (1, 2, \cdots, N)$. The variate value of interest for unit i is denoted by Y_i. A sample of size n selected from the population, and denoted by s, is an ordered sequence $i_1, \cdots, i_n$ ($i_j \in \mathbf{N}$; $j = 1, \cdots, n$, repetitions allowed) together with the associated sequence of observed variate values. A sample design is some countable set S of ordered sequences, s, together with a probability measure assigned by choosing a function $p(s) \geqq 0$ where $p(s)$ is the probability of selecting the sample s ($\sum_{s \in S} p(s) = 1$). It is commonly postulated that the principal objective is to use the sample data to make inferences about functions of $\mathbf{Y} = (Y_1, \cdots, Y_N)$; e.g., the finite population total, $\sum_{i=1}^{N} Y_i$.

Reference books treating this subject have appeared in clusters (in the early 1950's, early 1960's, late 1960's). Since the mid-1960's there have been a large number of significant new developments in the theory of sampling from finite populations. Thus, one looks forward to a new book treating this area. Ideally, such a book would unify at least some of the most important new ideas heretofore scattered throughout the literature; and would present the plethora of new results in a coherent, unified, notation. The tasks of unification and coherent presentation are particularly important in sample survey theory because some of the new developments represent significant departures from the foundations upon which most of the extant theory has been constructed.

The hallmarks of sample survey theory have been: (1) point estimation of parameters intended to "describe" the finite population (e.g., the finite population total); and (2) inferences based on the randomization distribution induced by the sample design. Recently, the extensive use of survey data for analytical purposes (i.e., to elucidate relationships among variables) has been recognized; and research has been initiated on the design and analysis of data from such surveys. Also, extensive research on the "foundations" of sampling from finite populations has been carried out. This has taken two forms: (1) delineation, support and extension of the randomization mode of inference; (2) development of alternative inferential methods. The latter commonly involve use of predictive inference for the nonsampled components of $\mathbf{Y}$ via superpopulation models linking

Received April 1976; revised July 1976.

$\mathbf{Y}$ to vectors of concomitant variables $\mathbf{X}_1, \cdots, \mathbf{X}_p$ where $\mathbf{X}_i = (X_{i1}, \cdots, X_{iN})$, $i = 1, 2, \cdots, p$, are generally assumed to be known. Prior distributions on the parameters of the superpopulation model are frequently added.

Publication of a book unifying the earlier results and the newer developments would be an invaluable aid to sample survey practitioners, researchers and teachers of university courses on sample survey theory and methodology. An important corollary would be stimulation of mathematical statisticians to initiate research on some of the important, challenging problems inherent in the design and analysis of samples from finite populations. While sample surveys are used to gather much data analyzed by statisticians, insufficient effort has been expended by mathematical statisticians to improve the theory. Moreover, development of new theory for sampling from finite populations should be appealing to some researchers because of the simplicity of the basic model.

Although *Statistical Theory of Sample Survey Design and Analysis* by H. S. Konijn is a useful addition to a statistician's library, it does not attain all of the goals cited above. Many of the important new directions in sample survey theory are not treated at all. In part, this is due to circumstances of publication. The book was published in 1973, but the latest reference I could find was dated 1970; and the majority of references are 1967 or earlier. In addition, it appears that the author's primary interest is reporting those new developments which proceed from the mainstream results developed earlier.

Not surprisingly, Konijn's book focuses on the design and analysis of descriptive surveys; i.e., those sample surveys whose end product is postulated to be a set of point estimates and estimates of precision. There is one chapter (of eight pages) treating "analytical" surveys; this accurately reflects the relative amounts of literature devoted to the two types of survey. As a survey practitioner located at a university, I find this situation unfortunate. In universities and smaller research organizations, surveys are conducted primarily for analytical purposes. However, procedures for the conduct of surveys focus largely on descriptive surveys. Thus, there is an important need to develop good survey designs where the objective is postulated to be the use of specific analytical approaches. Moreover, additional research is needed on methods of analysis of data from complex surveys when "standard" statistical techniques (e.g., contingency table analysis) are appropriate. It should be noted that surveys conducted by, for example, large governmental agencies, are also analytical in nature; i.e., the data will be used in an analytical manner. However, those who design and carry out such surveys are so remote from the (unknown) users that the design of such surveys generally cannot take such analytical uses into account.

Konijn's book is most effective as a reference volume for researchers in sample survey theory and methods. It is a faithful, very careful extension of the books on sample survey theory currently in print. Technical details are treated in a highly competent manner. I expect it to be less useful as a reference work for

most survey practitioners. The terse writing style may be advantageous for sophisticated readers, but will be disadvantageous for others. Unfortunately, important results and conditions for their application are sometimes difficult to locate. Given that there are very few illustrations of actual surveys, clarity of presentation of results would be enhanced by a theorem—proof—discussion format (e.g., the results in Section 4, pages 403–406). For use as a text, the instructor needs to supply both illustrations of surveys and exercises for the student to solve. Additional motivation for many topics is required. I found the manner of presentation to be cumbersome in some places and obtuse in others (see Section 4 below). More importantly, I would need to use supplementary material to cover topics which I find important, but which have been treated lightly or omitted altogether (see Section 3 below).

Konijn does cover several topics of considerable practical importance which are inadequately discussed in the literature, and uncommonly employed in practice. With emphasis on simple random sampling, he describes: (1) point estimation of the finite population cdf; and (2) confidence intervals for finite population quantiles and tolerance intervals. Superpopulation and infinite population concepts are treated in a chapter on the use of models in making inferences in sample survey problems. Use of these concepts in making inferences about finite population parameters is contrasted with the more customary method (i.e., based on the randomization distribution induced by the sample design).

Readers may wish to note the forthcoming book on probability sampling by J. Hájek; and the review article by Solomon and Zacks (1969) which covers new results in several areas.

2. Coverage. *Statistical Theory of Sample Survey Design and Analysis* is a book requiring knowledge of little more than elementary calculus. However, it is unlikely that a reader having only this minimal knowledge would possess the requisite manipulative skill. A first course in mathematical statistics would be a useful (but not mandatory) prerequisite.

The foundations of the book are described in Chapter 1. The finite population is defined and basic concepts (e.g., sample space, sample design, estimators) are described. Inference is assumed to proceed from the randomization distribution induced by the sample design adopted. Sampling distributions of estimators are treated briefly as are expectations of random variables. Processes to obtain a probability sample from a table of random digits are described with possibly excessive attention to detail.

Chapter 2 provides an extensive treatment of simple random sampling (SRS). In the derivation of the main results sampling with and without replacement are treated in parallel. I found the inclusion of simple cluster sampling (i.e., no subsampling) in this chapter to be useful; more complex sample designs can, thus, be discussed at an early stage of the book. Estimation of parameters of domains of study, and confidence intervals for the finite population mean, $\bar{Y}$,

are treated in detail. In the most novel section, confidence intervals for finite population quantiles and tolerance intervals are considered.

Assuming SRS, ratio (and related) estimators are discussed in Chapter 2A. Standard properties of the ratio estimator are derived, and common reduced-bias estimators are described. While regression, multiple regression, and difference estimators are treated in a cursory manner, the use of two-phase sampling in conjunction with ratio estimation is discussed in a careful manner.

Standard results for stratified simple random sampling are given in Chapter 3. These include use of ratio and difference estimators, estimation of parameters for domains of study, the collapsed strata method, use of two-phase sampling with stratification, post-stratification, and estimation of gains due to using stratified random sampling rather than SRS. There is an introduction to the use of replicated and balanced-half samples. Determination of the optimal number of strata, optimal strata boundaries and optimal choice of sample sizes in multivariate surveys are not discussed.

Considering sampling with replacement, estimators utilizing only the distinct units in the sample are treated carefully in Chapter 4. The "size" of sampling unit to select is discussed in Chapter 5. For example, given the data from a sample of "large" units, estimation of the precision which would have resulted from direct selection of a sample of "small" units (nested within the large units) is described.

A detailed treatment of sampling with unequal probabilities is presented in Chapter 6. Basic results assuming sampling both with and without replacement are outlined. What follows is a catalogue of methods to implement unequal probability sampling schemes. Surprisingly, one of the most commonly used methods—systematic sampling from the cumulated X_i—is not included in this chapter. (It is described in Chapter 9.) Unfortunately, there is minimal guidance about choice of methods.

Chapter 7 concerns cluster sampling with subsampling of the primary sampling units (p.s.u.'s) selected in the sample. Results are derived for standard situations such as: (1) SRS of both p.s.u.'s and second stage units (s.s.u.'s) with use of (a) mean per s.s.u., or (b) ratio estimators; (2) unequal probability sampling (with replacement) of p.s.u.'s, and several alternative methods of selecting s.s.u.'s; (3) unequal probability sampling of p.s.u.'s without replacement. Several comparisons of two-stage and single-stage sampling are made. Also presented is an extensive treatment of the optimal allocation of sample sizes in two-stage samples.

In Chapter 8, the potential use of stochastic models in survey sampling is considered. Comparisons between inferences based on (1) the customary randomization distribution, and (2) the use of stochastic models are made. Although the concept of a superpopulation is described and illustrated, the examples do not demonstrate the full potential of using this concept (see Section 3 below).

Konijn discusses systematic sampling in Chapter 9. The coverage is, principally, of well-known results. The customary comparisons among SRS, systematic

sampling and stratified random sampling are made. To secure estimates of variance the use of multiple random starts (or related methods) is recommended.

Chapter 10 contains an overview of nonsampling errors. Errors due to nonresponse and errors of measurement are discussed. Treated briefly are: (1) the effect (in SRS) of nonresponse on estimation of overall population parameters; (2) simple procedures for subsampling initially identified "nonrespondents;" and (3) adjustments to reflect differential probabilities of being "at-home" when interviews take place. Use of simple models to describe response errors permits evaluation of the contribution of such errors to the bias and variance of customary estimators. Estimation of such variances is also considered.

The use of repeated surveys to estimate both changes in parameters over time, and the values of parameters on the most recent occasion is described in Chapter 11. The material is illustrative as simplifying assumptions are made; however, the general theory is fully described in an Appendix.

In Chapter 12 there is a brief introduction to analytic uses of survey data. The discussion is limited to the comparison of the parameters of two populations, and focuses on the use of standardization to yield meaningful comparisons.

3. Additional topics. The coverage described in Section 2 above faithfully reflects the literature in sample survey theory and methodology. In this section of the review, several topics not covered or minimally treated in Konijn's book are discussed.[1] Most of these topics reflect relatively new and very important areas of inquiry. It is to be hoped that mathematical statisticians will be encouraged to initiate research on some of the important practical problems indicated below.

3.1. *Analytical uses of survey data.* Researchers extensively apply standard statistical techniques (e.g., regression analysis, contingency table analysis, factor analysis, principal component analysis) to data collected from complex sample surveys. Do these statistical techniques require modification? Preliminary indications suggest that modifications are necessary in some circumstances: see, for example, Porter (1973), Fuller (1973) for regression analysis; and Armitage (1966), Nathan (1972) and Sampford's discussion of Kish and Frankel (1974) for contingency table analysis. However, additional extensive research is required.

Observational studies are common in the social and health sciences, and merit separate consideration. They employ survey (and other) data, and there is considerable emphasis on special techniques (e.g., methods to remove bias). Two recent review papers may be consulted (McKinlay (1975), and Cochran and Rubin (1973)).

3.2. *Design of analytical sample surveys.* As noted in Section 1 of this review, many surveys have analytical objectives; they should be designed with such objectives in mind. While only preliminary research has been completed, some

[1] It should be recalled that some of these areas were inadequately developed at the time the book was written.

results are available in the literature. (Sedransk (1967) and Liao and Sedransk (1975) may be consulted for references.)

3.3. *Distributional properties of estimators.* The literature reflects a preoccupation with provision of point estimators, $\hat{\theta}$, and estimators of variance, $v(\hat{\theta})$. Presumably, one takes as a $(1 - \beta)100\%$ confidence interval for θ: $[\hat{\theta} - z_{\beta/2}\{v(\hat{\theta})\}^{\frac{1}{2}}, \hat{\theta} + z_{\beta/2}\{v(\hat{\theta})\}^{\frac{1}{2}}]$. Is this tenable in most surveys? Konijn gives the results of some empirical studies, but most of these relate to simpler sample designs than those met in practice. Moreover, the sample sizes considered appear to be much larger than those selected in "typical" surveys (see, e.g., Section 4.2, Section 7.3).

In stratified multistage surveys very few p.s.u.'s may be selected per stratum; it is conjectured that these small sample sizes are central in determining the distributional properties of the derived statistics. In such circumstances (and other ones) alternative confidence interval procedures may be required. Two promising methods are jackknifing and balanced repeated replication. (See Kish and Frankel (1974) for the results of some numerical trials.) Additionally, in many surveys (for example, of institutions) the distribution of Y is markedly skew. Here, one may profitably summarize data by estimating population quantiles. Konijn has described procedures appropriate for SRS. For stratified random sampling McCarthy (1965) gives useful results when proportional allocation is employed. An approximate procedure (of unverified efficacy) has been suggested by Woodruff (1952).

3.4. *Inferential framework for sampling from finite populations.* There is extensive research on the foundations of point estimation when sampling from finite populations. At the minimum, such results clarify the objectives of the sampling and estimation process and, therefore, are important for both practitioners and theoreticians. For example:

1. In what sense may estimators be said to have optimal properties? (See Godambe (1955, 1969), Godambe and Thompson (1973), Hartley and Rao (1968), Royall (1968).)

2. What is the "proper" role for the customary (randomization) inference in sampling from finite populations? (Konijn discusses this in Chapter 1, page 23 and pages 30–32; and in Chapter 8; one may also consult Godambe and Thompson (1973) and other papers referred to there.)

3. How does the customary inference relate to the use of conventional likelihood functions? (See Hartley and Rao (1968), Royall (1968).)

4. What role is there for Bayesian methods? (See Ericson (1969), Hartley and Rao (1968).)

3.5. *Bayesian methods.* Solutions to several important problems appear in the literature. For instance: (1) optimal allocation in stratified random sampling (Draper and Guttman (1968)); (2) optimal subsampling of nonrespondents (Ericson (1967)); (3) general principles of optimization (Rao and Ghangurde (1972)); (4)

estimation of finite population parameters in multistage surveys (Scott and Smith (1969)).

A general treatment of estimation of finite population parameters is given by Ericson (1969). Bayesian methods are now being applied in the solution of real survey problems (e.g., Sedransk (1975)). With further development of the theory and easier access to computers for applications, Bayesian ideas are likely to become more important in the conduct of surveys.

3.6. *Use of models.* Estimation of finite population totals when data on concomitant variables is available has been considered from a linear regression "superpopulation" (non-Bayesian) predictive approach. Optimal sampling strategies (i.e., choice of estimator and sample design) have been determined in specific situations (see, e.g., Royall (1970), Royall and Herson (1973)). Sampling methods which provide protection against misspecification of the superpopulation model have been proposed and evaluated (Royall and Herson (1973)).

3.7. *Problems inherent in handling large amounts of data.* Given a very large data set it is generally necessary to provide automated procedures for handling "outliers," for treating missing data, etc. Such problems are often exacerbated by the necessity of handling the data rapidly. Existing automated procedures need careful examination, and, where appropriate, more efficient techniques suggested.

4. Details. A careful reading of many sections of Konijn's book, and a cursory reading of the remainder suggests a careful attention to detail and a minimal number of technical errors and misstatements. However, a mathematical statistician, not an expert in sample survey theory who is asked to teach a course in this area, should carefully consider alternatives. Some of the following remarks are based on comments made by students in such a course. They were asked to read and comment critically upon Chapter 2 of Konijn's book after completion of a course oriented to Cochran's (1963) book.

(a) References to desired formulas are difficult as formulas are not numbered.

(b) Cumbersome proofs are sometimes supplied when more intuitive, easier proofs are available. For example, the initial proof that in SRS, $\text{Var}(\bar{y}) = S^2(N-n)/Nn$; basic proofs for systematic sampling when $N \neq nk$, page 361; confidence intervals for the Pth quantile of the finite population, pages 87–88.

(c) Footnotes are numerous and often contain significant technical details. They are difficult to read.

(d) In many places material is improperly sequenced. That is, a full understanding of the topic under discussion depends upon material to be covered in a subsequent section or chapter (e.g., page 20, Section 7; page 286, ll. 6–7; page 359, first footnote).

(e) There is a significant need for a glossary of terms: 1) there are undefined symbols, and quantities defined only in words; and 2) a mathematical expression

may be displayed, but the symbol used to denote it (e.g., S_i^{*2}, page 312) is defined in the *text*, not in the display.

(f) There are many typographical errors. Although most of these should be obvious to the careful reader, some formulas require careful reading. For instance, the extra N on page 298, l. 1 b; or the missing n in the denominator on page 301, l. 1 b would cause difficulty for the uncritical user.

(g) There are pedagogical problems exemplified by the following:

(i) Assuming SRS, the customary unbiased estimator of $S^2 = \sum_{i=1}^{N}(Y_i - \bar{Y})^2/(N-1)$ is written as $s^2 = \sum_{i=1}^{N} a_i(y_i - \bar{y})^2/(n-1)$ with the value of the random variable a_i depending on the presence ($a_i = 1$) or absence ($a_i = 0$) in the sample of the ith population unit. My experience suggests this to be unwise; novices will often write $E(s^2) = \sum_{i=1}^{N}(y_i - \bar{y})^2\{E(a_i)\}/(n-1)$ with $E(a_i) = n/N$ without realizing that $\bar{y} = \sum_{i=1}^{N} a_i Y_i/n$. This type of expression is employed throughout.

(ii) Treatment of simple cases in a complex manner: The end product of Chapter 5, Section 1 is an Anova for the special case where n clusters are selected by SRS from N clusters, and all $\bar{M}$ units in each selected cluster are enumerated. However, this section is extraordinarily difficult to follow because of (1) the unnecessary introduction of notation for unequal sized clusters and two stage sampling; and (2) the orientation of the section to the special case.

While it is infeasible to supply an exhaustive list of the sections in Konijn's book not treated optimally, a few topics of potential import for researchers are noted below.

(a) Rao (1975) has shown that there is an error in Konijn's expression (page 308) for the variance estimator corresponding to Wilks' procedure for two-stage sampling with p.s.u.'s included with unequal probabilities.

(b) The use of two-phase sampling to improve estimation of domain parameters is considered in Chapter 3, Section 9.7. Procedures which are more satisfactory have been suggested by deGraft-Johnson and Sedransk (1973). (Also, see the references therein.)

(c) The discussion of two-phase sampling for stratification (Sections 9.3, 9.4) is unsatisfactory. See Rao (1973) for an alternative method.

(d) The choice of cost functions for illustrative purposes is important, and their limitations should be noted. For instance, in choosing the appropriate size of unit (Chapter 5, Section 2), the cost functions may not adequately represent the differential travel costs which are often important in such surveys.

(e) With extensive use of asymptotic methods to derive approximations for the moments of ratio estimators, etc., a theorem to permit easy identification of the terms (of specified order) to be retained should be quoted. Inclusion of such a theorem would prevent an inclarity such as the one in the second footnote on page 99.

(f) For an estimator, $\hat{\theta}$, of θ, $\mathrm{Var}(\hat{\theta}) = E(\hat{\theta}^2) - \{E(\hat{\theta})\}^2$; and $\hat{\theta}^2 - \hat{\tau}$ is frequently used as an (unbiased) estimator of $\mathrm{Var}(\hat{\theta})$ where $E(\hat{\tau}) = \{E(\hat{\theta})\}^2$. It is to be hoped

that research will yield more appropriate ways of estimating variances (cf. Sections 3.4, 3.5 and 3.6 above).

REFERENCES

[1] ARMITAGE, P. (1966). The chi-square test for heterogeneity of proportions after adjustment for stratification. *J. Roy. Statist. Soc. Ser. B* **28** 150–163.

[2] COCHRAN, W. G. (1963). *Sampling Techniques*, 2nd. ed. Wiley, New York.

[3] COCHRAN, W. G. and RUBIN, D. (1973). Controlling bias in observational studies: a review. *Sankhyā Ser. A* **35** 417–446.

[4] DEGRAFT-JOHNSON, K. T. and SEDRANSK, J. (1973). Estimation of domain means using two-phase sampling. *Biometrika* **60** 387–393.

[5] DRAPER, N. and GUTTMAN, I. (1968). Some Bayesian stratified two-phase sampling results. *Biometrika* **55** 131–140.

[6] ERICSON, W. (1967). Optimal sampling design with nonresponse. *J. Amer. Statist. Assoc.* **62** 63–78.

[7] ERICSON, W. (1969). Subjective Bayesian models in sampling finite populations (with discussion). *J. Roy. Statist. Soc. Ser. B* **31** 195–233.

[8] FULLER, W. (1973). Regression analysis for sample surveys. Technical Report, Statistical Laboratory, Iowa State Univ.

[9] GODAMBE, V. P. (1955). A unified theory of sampling from finite populations. *J. Roy. Statist. Soc. Ser. B* **17** 268–278.

[10] GODAMBE, V. P. (1969). Some aspects of the theoretical developments in survey-sampling. *New Developments in Survey Sampling* (N. L. Johnson and H. Smith, eds.) 27–58. Wiley-Interscience, New York.

[11] GODAMBE, V. P. and THOMPSON, M. E. (1973). Philosophy of survey-sampling practice. Technical Report, Statistics Department, Univ. of Waterloo.

[12] HARTLEY, H. O. and RAO, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika* **55** 547–558.

[13] KISH, L. and FRANKEL, M. (1974). Inference from complex samples (with discussion). *J. Roy. Statist. Soc. Ser. B* **36** 1–37.

[14] LIAO, H. and SEDRANSK, J. (1975). Sequential sampling for the comparison of domain means. *Biometrika* **62** 690–693.

[15] MCCARTHY, P. J. (1965). Stratified sampling and distribution-free confidence intervals for a median. *J. Amer. Statist. Assoc.* **60** 772–783.

[16] MCKINLAY, S. (1975). The design and analysis of the observational study—a review. *J. Amer. Statist. Assoc.* **70** 503–520.

[17] NATHAN, G. (1972). On the asymptotic power of tests for independence in contingency tables from stratified samples. *J. Amer. Statist. Assoc.* **67** 917–920.

[18] PORTER, R. (1973). On the use of survey sample weights in the linear model. *Ann. of Econ. and Social Measurement* **2** 141–158.

[19] RAO, J. N. K. (1973). On double sampling for stratification and analytical surveys. *Biometrika* **60** 125–134.

[20] RAO, J. N. K. (1975). Unbiased variance estimation for multistage designs. Technical Report, Department of Mathematics, Carleton Univ.

[21] RAO, J. N. K. and GHANGURDE, P. (1972). Bayesian optimization in sampling finite populations. *J. Amer. Statist. Assoc.* **67** 439–443.

[22] ROYALL, R. (1968). An old approach to finite population sampling theory. *J. Amer. Statist. Assoc.* **63** 1269–1279.

[23] ROYALL, R. (1970). On finite population sampling theory under certain linear regression models. *Biometrika* **57** 377–388.

[24] ROYALL, R. and HERSON, J. (1973). Robust estimation in finite populations, I. *J. Amer. Statist. Assoc.* **68** 880–889.

[25] SCOTT, A. and SMITH, T. M. F. (1969). Estimation in multi-stage surveys. *J. Amer. Statist. Assoc.* **64** 830–840.

[26] SEDRANSK, J. (1967). Designing some multi-factor analytical studies. *J. Amer. Statist. Assoc.* **62** 1121–1140.

[27] SEDRANSK, J. (1975). Sampling problems in the estimation of the money supply. Technical Report No. 29, Statistical Science Division, State Univ. of New York at Buffalo.

[28] SOLOMON, H. and ZACKS, S. (1969). Optimal design of sampling from finite populations. *J. Amer. Statist. Assoc.* **65** 653–677.

[29] WOODRUFF, R. (1952). Confidence intervals for medians and other position measures. *J. Amer. Statist. Assoc.* **47** 635–646.

STATISTICAL SCIENCE DIVISION
SUNY AT BUFFALO
4230 RIDGE LEA ROAD
AMHERST, NEW YORK 14226