Werk

Titel: On UMV-Estimators in Survey Sampling.

Autor: Lanke, J.

Jahr: 1973

PURL: https://resolver.sub.uni-goettingen.de/purl?358794056_0020|log59

Kontakt/Contact

Digizeitschriften e.V.
SUB Göttingen
Platz der Göttinger Sieben 1
37073 Göttingen

info@digizeitschriften.de

On UMV-Estimators in Survey Sampling

By J. LANKE, LUND[1])

Zusammenfassung: Es wird die Existenz von Schätzungen mit gleichmäßig minimaler Varianz in der Klasse aller linearen erwartungstreuen Schätzungen des Populationstotales besprochen. Einige Bemerkungen werden zum Begriff "necessary bestness" gemacht.

Summary: The existence of uniformly minimal variance estimators in the class of all linear unbiased estimators of the population total is discussed. Some remarks are given on the concept of "necessary bestness".

1. Introduction

Suppose a sampling survey has been performed and suppose we are requested to estimate a certain numerical property of the population, e.g. the population total of the characteristic under study. Then we have many possible estimators to choose among and even restriction to unbiased estimators does usually not single out a unique estimator. In such a case it seems natural to apply some optimality criterion. If we agree to use the variance as risk function, the strongest possible optimality property of course is to have uniformly minimal variance. It appears to be widely accepted that UMV-estimators do in general not exist in the class of linear estimators where "in general" means "for sampling designs which are not unicluster designs". To the best of our knowledge no *correct* proof of this fact has, however, yet been published and we give what we hope to be such a proof. Further we give some comments on the concept of "necessary bestness" and point out an error in a recent paper.

2. Notations and definitions

The set $U = \{1, 2, \ldots, N\}$ is called the population and its elements are called units; a subset s of U is called a sample and S denotes the set of all the 2^N possible samples; a (sampling) design is a non-negative function p on S such that $\Sigma_s p(s) = 1$;

[1]) JAN LANKE, Dept. of Math. Stat., Box 725, S-22007 Lund, Sweden.

the inclusion probabilities $\pi_1, \pi_2, \ldots, \pi_N$ of a design p are defined by $\pi_i = \Sigma_{s \ni i} p(s)$ where the sum is over all samples s which contain the unit i; a design p is called a unicluster design if every unit i occurs in at most one sample s with $p(s) > 0$, i.e. if $p(s_1) > 0$, $p(s_2) > 0$, $s_1 \neq s_2$ implies $s_1 \cap s_2 = \emptyset$; the character of interest is a real function y defined on U and $y(i)$ is denoted y_i; the vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$ is called the parameter and is considered as an element in the parameter space R_N; any function on R_N is called a parametric function and the particular parametric function $T(\boldsymbol{y}) = \Sigma_i y_i$ is called the population total; an estimator is a function e on $S \times R_N$ such that $e(s, \boldsymbol{y})$ depends upon $\boldsymbol{y}$ only through those y_i for which $i \in s$, i.e. such that for every $s \in S$ an every $i \in U$ the function $y_i \to e(s, \boldsymbol{y})$ is constant if $i \notin s$; an estimator e is linear if, for every $s \in S$, the function $\boldsymbol{y} \to e(s, \boldsymbol{y})$ is linear, i.e. if $e(s, \boldsymbol{y}) = \Sigma_i \beta(s, i) y_i$ where of course $\beta(s, i) = 0$ if $s \not\ni i$; to every design p with $\pi_i > 0$ for all i corresponds a certain estimator, the Horvitz-Thompson estimator, defined by $e_{HT}(s, \boldsymbol{y}) = \Sigma_{i \in s} y_i / \pi_i$; the expectation and the variance of an estimator e in a design p are defined by $E_p(e, \boldsymbol{y}) = \Sigma_s p(s)\, e(s, \boldsymbol{y})$ and $V_p(e, \boldsymbol{y}) = E_p(e^2, \boldsymbol{y}) - E_p^2(e, \boldsymbol{y})$; the estimator e is said to be unbiased for the parametric function F (in the design p) if $E_p(e, \boldsymbol{y}) = F(\boldsymbol{y})$ for all $\boldsymbol{y} \in R_N$; when no confusion can occur we will drop the subscripts in E_p and V_p; an estimator e in a certain class C of unbiased estimators of a certain parametric function is admissible in C if there no e' in C such that $V(e', \boldsymbol{y}) \leq V(e, \boldsymbol{y})$ for all $\boldsymbol{y} \in R_N$ with strict inequality for at least one $\boldsymbol{y}$.

3. Some lemmas and an example

Our first lemma is very well known and easily established. We record it here only for completeness.

Lemma 1. A sampling design p admits a linear unbiased estimator of $T(\boldsymbol{y})$ if and only if it has $\pi_i > 0$ for all i.

Although it will be of no importance to us, let us remark that Lemma 1 remains true also if we delete the word "linear".

The following result is implicit in GODAMBE [1960] but appears not to be as widely known as it deserves.

Lemma 2. For any sampling design p which has $\pi_i > 0$ for all i, the Horvitz-Thompson estimator e_{HT} has uniformly minimal variance in the class of all linear unbiased estimators of $T(\boldsymbol{y})$ when the parameter $\boldsymbol{y}$ is restricted to the set $R_N^{(1)}$ of vectors in R_N having at most one non-zero component.

Our next lemma may well be known to some research workers but the author has not been able to find it explicitly stated anywhere.

Lemma 3. If p is a non-unicluster design, the Horvitz-Thompson estimator e_{HT} does *not* have uniformly minimal variance in the class of all linear unbiased estimators of $T(\boldsymbol{y})$ when the parameter $\boldsymbol{y}$ is restricted to the set $R_N^{(2)}$ of vectors in R_N having at most two non-zero components.

Proof. Since p is non-unicluster there exist two samples s_1 and s_2 which both have positive probability and two units, say 1 and 2, such that

$$1 \in s_1 \qquad 1 \in s_2 \qquad 2 \in s_1 \qquad 2 \notin s_2 \,.$$

The estimator

$$e(s, \boldsymbol{y}) = \begin{cases} e_{HT}(s_1, \boldsymbol{y}) - y_1 p(s_2) & \text{for} \quad s = s_1 \\ e_{HT}(s_2, \boldsymbol{y}) + y_1 p(s_1) & \text{for} \quad s = s_2 \\ e_{HT}(s, \boldsymbol{y}) & \text{else} \end{cases}$$

is obviously linear and is easily seen to be unbiased provided p has $\pi_i > 0$ for all i; if $\pi_i = 0$ for some i, the lemma is trivial. A simple calculation shows that if $y_k = 0$ for $3 \leq k \leq N$, we have

$$E(e_{HT}^2 - e^2, \boldsymbol{y}) = A y_1^2 + B y_1 y_2$$

with certain constants A and B where $B \neq 0$. If we choose y_1 small, $|y_2|$ large and the sign of y_2 conveniently we see that $V(e_{HT}, \boldsymbol{y})$ cannot be smaller than $V(e, \boldsymbol{y})$ uniformly when $\boldsymbol{y} \in R_N^{(2)}$. QED

The following simple observation will prove useful for the study of unicluster designs.

Lemma 4. If p is a unicluster design with $\pi_i > 0$ for all i, then e_{HT} is the only linear unbiased estimator of $T(\boldsymbol{y})$.

Proof. The fact that e_{HT} is unbiased for $T(\boldsymbol{y})$ under the condition "$\pi_i > 0$ for all i" follows directly from the definition. To prove the uniqueness, suppose $e(s, \boldsymbol{y}) = \Sigma_i \beta(s, i) y_i$ is unbiased for $T(\boldsymbol{y})$. Since each unit i in U belongs to exactly one of those samples $s_1, s_2, \ldots, s_M$ which have positive probability under p, it follows that each y_i occurs exactly once in the left hand side of the identity

$$\sum_{k=1}^{M} p(s_k) \sum_{i \in s_k} \beta(s_k, i) y_i = \sum_{i=1}^{N} y_i \,.$$

Thus $\beta(s_k, i) = 1/p(s_k)$ and the uniqueness follows. QED

The definition of an estimator, as given in the previous section, requires $e(s, \boldsymbol{y})$ to be a function of only those y_i for which $i \in s$. This naturally does not rule out the possibility that there is a sample s and a unit $i \in s$ such that $e(s, \boldsymbol{y})$ does not depend upon y_i; after all, a constant function is also a function! In case an estimator has this property we will say that it does not depend upon $\boldsymbol{y}$ fully. Now it might be reasonable to guess that such an estimator cannot be a very good estimator. However, the following example shows that it can be admissible.

Example. Let $U = \{1, 2, 3\}$, let p assign the probability 1/3 to each of samples $\{1\}$, $\{2\}$, $\{2, 3\}$ and let e be given by

$$e(s, \boldsymbol{y}) = \begin{cases} 3y_1 & \text{for} \quad s = \{1\} \\ 3y_2 & \text{for} \quad s = \{2\} \\ 3y_3 & \text{for} \quad s = \{2, 3\} . \end{cases}$$

Since $e(\{2, 3\}, \boldsymbol{y})$ does not depend upon y_2, this estimator does not depend upon $\boldsymbol{y}$ fully. The fact that e is unbiased for $T(\boldsymbol{y})$ is obvious. To show that e is admissible in the class of all unbiased estimators of $T(\boldsymbol{y})$, we must show that, given any unbiased e' of $T(\boldsymbol{y})$ with

$$V(e', \boldsymbol{y}) \leq V(e, \boldsymbol{y}) \quad \text{for all } \boldsymbol{y} , \tag{1}$$

it follows that

$$V(e', \boldsymbol{y}) = V(e, \boldsymbol{y}) \quad \text{for all } \boldsymbol{y} .$$

But every estimator can be written

$$e'(s, \boldsymbol{y}) = \begin{cases} f_1(y_1) & \text{for} \quad s = \{1\} \\ f_2(y_2) & \text{for} \quad s = \{2\} \\ f_3(y_2, y_3) & \text{for} \quad s = \{2, 3\} . \end{cases}$$

Since $V(e, \mathbf{0}) = 0$, (1) implies that $V(e', \mathbf{0}) = 0$ which gives $f_1(0) = f_2(0) = f_3(0, 0) = 0$. Now $E(e', \boldsymbol{y}) = T(\boldsymbol{y})$ gives

$$f_1(y_1) = 3y_1 , \qquad f_3(y_2, y_3) = 3y_2 + 3y_3 - f_2(y_2) .$$

Finally (1) gives

$$3y_3(3y_2 - f_2(y_2)) \leq 3y_2 f_2(y_2) - f_2^2(y_2)$$

which can be true for all y_3 only if $f_2(y_2) = 3y_2$, i.e. if $e' = e$. Thus no unbiased estimator e' of $T(\boldsymbol{y})$ can be strictly better than e.

4. On UMVLU-estimators

In this section we are going to discuss the following theorem.

Theorem. A sampling design p admits a uniformly minimal variance estimator in the class of all linear unbiased estimators of $T(\boldsymbol{y})$ (a UMVLUE of $T(\boldsymbol{y})$, for short) if and only if it is a unicluster design with $\pi_i > 0$ for all i.

When sketching the history of this important theorem we will use "design" as short for "design with $\pi_i > 0$ for all i", thereby following what most previous authors, perhaps unconsciously, have done.

GODAMBE [1955] formulates and claims to prove the theorem that *no* sampling design admits a UMVLUE of $T(\boldsymbol{y})$; his proof, however, contains a slip, a consequence of which is that he does not observe that the unicluster designs form an exception.

GODAMBE's theorem is reproduced in KOOP [1963] with a new proof. Further KOOP adds the comment that it is well known that a UMVLUE of $T(\boldsymbol{y})$ exists for some sampling designs, e. g. for simple random sampling without replacement; the rationale behind this comment has escaped the present author. That KOOP's proof is fundamentally wrong is easily seen from the fact that, as KOOP himself remarks, the same proof shows the non-existence of a UMVE of $T(\boldsymbol{y})$ not only in the whole class of all linear unbiased estimators but also in certain other sets, in fact in any set, of linear unbiased estimators; now any set consisting of just one linear unbiased estimator furnishes a counter-example to this alleged result.

HEGE [1965] first quotes GODAMBE's result, then claims to quote a remark of GODAMBE that some exceptions exist and finally proves the existence of "some more" exceptions. Her quotation from GODAMBE on exceptions, however, results from a complete misunderstanding of what GODAMBE writes. The exceptions she herself points out are the unicluster designs. It should be observed that she neither proves nor claims to prove the converse, i.e. that the unicluster designs are the only exceptions. (The statement to the contrary in HEGE [1966] appears to result from a slip: she writes "only if" when she means "if".)

HANURAV [1965] gives the correct formulation of the theorem, but his proof of the "only if"-part contains an error.

GODAMBE [1965] states that a sampling design admits no UMVLUE except if it is a unicluster design "which is a trivial case" but no proof is given.

HANURAV [1966], quoting GODAMBE [1965], gives the correct form of the theorem (but curiously enough fails to give GODAMBE full credit for the correct formulation). He then gives a proof which differs from his above-mentioned one. The present authors regrets that at a certain stage of this second proof he is unable to follow the argument. In fact, it seems likely that HANURÁV is under the impression that an estimator e must be inadmissible if it does not depend upon $\boldsymbol{y}$ fully, i. e. if there is a sample s_0 with $p(s_0)>0$ and a unit $i \in s_0$ such that the function $y_i \to e(s_0, \boldsymbol{y})$ is constant. This is, however, not the case as is seen from our example in section 3. HANURAV's proof can be saved if we use the following observation which is due to RAMAKRISHNAN [1972]: If an estimator is a UMVLUE for $T(\boldsymbol{y})$, then it must depend upon $\boldsymbol{y}$ fully. (Hint of proof: If e does not depend upon $\boldsymbol{y}$ fully, a linear function with zero expectation can be constructed in such a way that it is correlated with e. But then e cannot be a UMVLUE for $T(\boldsymbol{y})$.)

MURTHY and SINGH [1969] observe that if a class C contains an admissible estimator e as well as an estimator e' which for at least one $\boldsymbol{y}$ is better than e, then C contains no UMVE. If C is the class of LUE's of $T(\boldsymbol{y})$ it follows from GODAMBE [1960] that e_{HT} can serve as e; as an argument for the existence of an e' with the desired property MURTHY and SINGH refer to MURTHY [1957] and to RAO and BAYLESS [1969], two references which are easily seen to be equally irrelevant.

In connection with linear sufficiency RAMAKRISHNAN [1970] gives a theorem

which, when combined his remark above, furnishes a proof of our theorem.

To summarize: The correct formulation of the theorem was given independently by GODAMBE [1965] and HANURAV [1965]. As far as the present author can see, no correct proof has yet been published. Now we intend to give a short (and, as we hope, correct) proof.

Proof. Lemma 1 shows that for a design p with $\pi_i = 0$ for some i, no LUE exists and thus no UMVLUE can exist for such a design. For a design p with $\pi_i > 0$ for all i, Lemma 2 shows that if p admits a UMVLUE e, then we must have $e = e_{HT}$. Now Lemma 3 tells us that if the design is a non-unicluster design, then e_{HT} is not a UMVLUE and hence no UMVLUE exists in this case. On the other hand, if the design is a unicluster design with $\pi_i > 0$ for all i, Lemma 4 shows that there is exactly one LUE, which then of course is a UMVLUE. QED

5. On the concept of necessary bestness

When discussing linear estimators, PRABHU AJGAONKAR [1965] introduces the concept of necessary bestness and later [1969] generalizes it to necessary bestness of order k where the case $k = 1$ corresponds to the previously defined concept. A condensed version of his definition, which has the advantage of making sense also for non-linear estimators, is the following.

Definition. An estimator e in a class C of unbiased estimators of a certain parametric function is said to be a necessary best estimator of order k in C if it is a UMVE in C when $\boldsymbol{y}$ is restricted to the part $R_N^{(k)}$ of R_N where at most k coordinates are non-zero.

The author is informed that a similarly modified definition also occurs in a technical report by RAO and SINGH but he has not had the opportunity to see this report.

PRABHU AJGAONKAR [1969] states and claims to prove

(i) e_{HT} is a NBE of order 1 in L

(ii) e_{HT} is a NBE of order 2 in L

(iii) there is no NBE of order 3 in L

where NBE means necessary best estimator and where L denotes the class of all linear unbiased estimators of $T(\boldsymbol{y})$. The first of these statements, which is known from HEGE [1967] and also is discussed by OGUS [1969], is an immediate consequence of our Lemma 2 in section 3 if the design has $\pi_i > 0$ for all i; otherwise it is false. The fact that statement (ii) is wrong (unless the design is a unicluster design with $\pi_i > 0$ for all i) is immediately seen from our Lemma 3; in fact (i) and Lemma 3 prove that, save the exception just mentioned, we have

(ii') there is no NBE of order 2 in L.

Finally, the truth of (iii) (except for unicluster designs with $\pi_i > 0$ for all i) of course follows directly from (ii').

References

GODAMBE, V. P.: A unified theory of sampling from finite populations. J. Roy. Statist. Soc. Ser. B, **17,** 269–278, 1955.

GODAMBE, V. P.: An admissible estimate for any sampling design. Sankhyā, **22,** 285–288, 1960.

GODAMBE, V. P.: A review of the contributions towards a unified theory of sampling from finite populations. Rev. Inst. Internat. Statist., **33,** 242–258, 1965.

HANURAV, T. V.: Optimum sampling strategies. Ph. D. thesis submitted to the Indian Statistical Institute, 1965.

HANURAV, T. V.: Some aspects of unified sampling theory. Sankhyā, A, **28,** 175–204, 1966.

HEGE, V.: Sampling designs which admit uniformly minimum variance unbiased estimate. Calcutta Statist. Assoc. Bull., **14,** 160–162, 1965.

HEGE, V.: Admissible estimates for any sampling design. Calcutta Statist. Assoc. Bull., **15,** 120–126, 1966.

HEGE, V.: An optimum property of the Horvitz-Thomson [!] estimate. J. Amer. Statist., Assoc., **62,** 1013–1017, 1967.

KOOP, J. C.: On the axioms of sample formation and their bearing on the construction of linear estimators in sampling theory for finite universes. Metrika, **7,** 81–114 and 165–204, 1963.

MURTHY, M. N.: Ordered and unordered estimators in sampling without replacement. Sankhyā, **18,** 379–390, 1957.

MURTHY, M. N., and M. P. SINGH: On the concepts of best and admissible estimators in sampling theory. Sankhyā, A, **31,** 343–354, 1969.

OGUS, J.: A note on the "necessary best estimator". J. Amer. Statist. Assoc., **64,** 1350–1352, 1969.

PRABHU AJGAONKAR, S. G.: On a class of linear estimators in sampling with varying probabilities without replacement. J. Amer. Statist. Assoc., **60,** 637–642, 1965.

PRABHU AJGAONKAR, S. G.: On the non-existence of a best estimator for the entire class of linear estimators. Sankhyā, A, **31,** 455–462, 1969.

RAMAKRISHNAN, M. K.: Optimum estimators and strategies in survey sampling. Ph. D. thesis submitted to the Indian Statistical Institute, 1970.

RAMAKRISHNAN, M. K.: Personal communication, 1972.

RAO, J. N. K., and D. L. BAYLESS: An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of units per stratum. J. Amer. Statist. Assoc., **64,** 540–559, 1969.

Received April 25, 1972