STOCKHOLMS UNIVERSITET HT2022

Statistiska institutionen
Dan Hedlin

Sample surveys, ST306G

Examination 2022-10-26, 14.00 – 19.00

Approved aids:

1. Pocket calculator
2. Language dictionary

Separate pages with notes are not allowed.

The exam comprises 9 items, numbered 1 to 9. The maximum number of points is 50. 25 points will give you at least grade E. To obtain the maximum number of points full and clear motivations are required unless otherwise stated. You may write in English or Swedish. **There are some pages at the end of the exam with formulae that you may wish to use.**

In each of the five questions below one of the items a, b, c, d or e is incorrect. Which one? For each of the questions 1-5, answer with only one letter, a-e. Motivation is not required. Maximum 10 points (2 for each of 1-5).

1.

a) A domain of study, or domain for short, is a subset of the target population.
b) Two domains in the same survey must not overlap.
c) Typical domains in a labour force survey are, for example, age groups.
d) In many surveys you are as (or even more) interested in estimates for domains than in estimates for the whole population.
e) One common aim of stratification is to (try to) create strata that are similar to important domains.

2.

a) Suppose the population consists of pupils in all schools in a region. The set of samples that can be drawn with one-stage cluster sampling of schools is a subset of the set of samples that can be drawn with simple random sampling of pupils.
b) We have focused on design-based inference in the course. Which of the possible samples that is drawn is viewed as random in design-based inference and this randomness plays an integral part in the definition of variance.
c) It is not incorrect to treat n_d (the number of sampling units in the part of the sample that falls in domain d) in poststratification as non-random, although it is random.
d) If the inclusion probability is n/N (sample size divided by population size) for all units in the population, then the sampling design is simple random sampling.
e) Three choices to make when designing a stratified sample is
    - how to define strata

- what sampling design or sampling designs to use
- how to allocate the sample to strata

3.

a) The main difference between a census and a sample survey is that in a census you try to collect data from all units in the population rather than from a sample of units.

b) The main difference between register-based statistics and censuses is that in a census the statistician has control over what data to collect, in register-based statistics the statistician has to rely on the register, which usually has some administrative purpose rather than a statistical purpose.

c) In sample surveys, the sample size needs to be at least 40 to be scientifically valid.

d) While a census does not involve sampling, a census does not eliminate statistical uncertainty. It may still suffer from, for example, nonresponse.

e) While register-based statistics does not involve sampling, a register does not eliminate statistical uncertainty. It may still suffer from, for example, reporting delays.

4.

a) One way to understand why it is not possible to construct an unbiased estimator of the variance for a systematic sampling design is to view a systematic sample as a cluster sample with only one sampled cluster.

b) Both strata in stratified simple random sampling and clusters in one-stage cluster sampling are subsets of the population such that each observation unit belongs to one and only one subset.

c) Suppose you compare the variance for the estimation of the population mean for two sampling designs: simple random sampling and stratified simple random sampling with proportional allocation. It happens fairly often in practice that simple random sampling is better than stratified simple random sampling with proportional allocation in terms of precision.

d) What width of a confidence interval that is needed should be decided by the user of the statistics (although the user may need some help).

e) If a client ask you, the statistician, ‘what sample size do I need for my survey’, two important questions to ask the client are: 1) are you planning to produce estimates for any domains 2) will your survey contain continuous variables (such as income).

5.

a) The unbiased estimator of the variance for the Horvitz-Thompson estimator of the total for simple random sampling is $\hat{V}(\hat{t}_y) = N^2\left(1-\frac{n}{N}\right)\frac{s_y^2}{n}$. Another unbiased estimator in the same situation is $\hat{V}(\hat{t}_y) = N^2\frac{s_y^2}{n}$.

b) If the samples in stratified simple random sampling are dependent (that is, if $P(s_h, s_k) \neq P(s_h)P(s_k)$ for samples s_h and s_k in stratum h and stratum k), then the variance of an estimate of the population mean is not $V(\bar{y}_{str}) = \sum_{h=1}^{H}\frac{N_h^2}{N^2}\left(1-\frac{n_h}{N_h}\right)\frac{S_h^2}{n_h}$

c) Suppose that the study variable y is binary and that the sampling design is simple random sampling. The following argument shows that the estimator of a proportion is a special case of the estimator of the mean: $\bar{y}_s = \frac{1}{n}\sum_{i\in s} y_i = \frac{1}{n}\tau_s = \hat{p}$, where τ_s is the number of sample units for which $y_i = 1$ and $\hat{p}$ is an estimator of the population proportion.

d) Suppose $Z_i = 1$ if sample unit i is included in the sample and $Z_i = 0$ if unit i is not included in the sample. Then $E(Z_i)$ is the inclusion probability of the sample unit i.

e) In optimal allocation (aka Neyman allocation) the stratum sample sizes n_h are determined so that they are proportional to $N_h S_h$ (see the formulas at the end of this exam paper for notation). If the purpose is not to minimise $V(\hat{t}_y)$ then n_h should not necessarily be proportional to $N_h S_h$.

6.

A population consists of nine units. The units, labelled A, B, …, I, and their values of the study variable y are shown in the table below.

Label	y
A	2
B	2
C	3
D	4
E	4
F	4
G	5
H	5
I	6
Sum	35

The possible samples with the sampling design that the statistician has devised are
$s_1 = \{\text{A,B,C}\}$, $s_2 = \{\text{A,B,D}\}$, $s_3 = \{\text{A,C,D}\}$, $s_4 = \{\text{B,C,D}\}$, $s_5 = \{\text{B, C, E}\}$,
$s_6 = \{\text{B, C, F}\}$, $s_7 = \{\text{E, F, G}\}$, $s_8 = \{\text{E, G, H}\}$, $s_9 = \{\text{F, G, I}\}$, $s_{10} = \{\text{G, H, I}\}$
and the probability of drawing any one of them is 1/10.

a) What is the inclusion probability π_A, that is, the inclusion probability of unit A?

b) Let the Horvitz-Thompson estimate based on sample s_j, $j = 1, 2, \dots, 10$, be denoted by $\hat{t}_{s_j}$. What is the expected value of the Horvitz-Thompson estimator of the total, that is, what is

$$E(\hat{t}_y) = \Pr(s_1)\,\hat{t}_{s_1} + \Pr(s_2)\,\hat{t}_{s_2} + \Pr(s_3)\,\hat{t}_{s_3} + \Pr(s_4)\,\hat{t}_{s_4} + \Pr(s_5)\,\hat{t}_{s_5} + \Pr(s_6)\,\hat{t}_{s_6} + \Pr(s_7)\,\hat{t}_{s_7} + \Pr(s_8)\,\hat{t}_{s_8} + \Pr(s_9)\,\hat{t}_{s_9} + \Pr(s_{10})\,\hat{t}_{s_{10}}?$$

c) Does this sampling design have a specific name; if so, what is this sampling design called? No motivation is required.

d) Suppose now that a sample of size $n = 3$ is drawn from the population with the systematic sampling design. The first unit to be included in the sample was determined by a random selection of one of the units A, B and C, each with probability 1/3. What is the population sum of the inclusion probabilities, that is, $\pi_A + \pi_B + \pi_C + \pi_D + \pi_E + \pi_F + \pi_G + \pi_H + \pi_I$?

Now we turn to a different population with $N = 5$ units. The inclusion probabilities are as in the table below. There is an auxiliary variable x which is known to be strongly correlated with the study variable y. The values of x are known for every unit in the population prior to sampling. The sample size is 3. A sample is drawn with the inclusion probabilities in the table below and it turns out that the sample is $\{$A,B,D$\}$. The values of the study variable $y_A = 1.1$, $y_B = 3.1$ and $y_D = 145$ are observed.

e) Suggest a better sampling design for estimating the total, still with sample size 3.

Label	x	Inclusion probability
A	1	0.6
B	2	0.6
C	1	0.5
D	3	0.7
E	130	0.6

Maximum 14 points.

7.

a) Give precise verbal definitions of: domain, stratum and poststratum.

b) In January 2022 Renault in Sweden took an SRS (simple random sample) of owners of a Renault car (according to the Swedish car register) to obtain data on customer satisfaction. The sample size was 2000. 1000 responded and we assume MCAR (missing completely at random). As the overall customer satisfaction was disappointingly low, the company wanted to study customer satisfaction of those Renault owners who had a car made in 2021. Estimate the proportion of owners of a Renault made in 2021 who were pleased with their Renault car. Estimate also the variance for the estimated proportion. There were 200 000 owners of a Renault car in January 2022. 200 of those had a Renault made in 2021. You can use $s_{yd}^2 = 0.10$ if you do not know how to compute it.

c) Give two reasons why MCAR may be a poor assumption in this survey.

Made in year	Number of respondents	Number who were pleased with their Renualt
2021	76	72
2020	139	88
2019	166	92
2018 or before 2018	619	201
Sum	1000	453

Maximum 9 points.

8.

A simple random sample was taken to estimate the prevalence of covid-19 in a population. Interviewers made contact with the selected people and asked if they could take a nasal swab test. In an area with 20 000 people, an SRS of 465 people was taken. Interviewers were able to collect swab tests from 200 people. 25 of those had a positive test. Assume that a positive test result means covid and that a negative results means that the tested person does not have covid. The row for 'whole area' in tables 1 and 2 below contain some data.

a) State one likely reason for nonresponse in this survey, and discuss briefly the likely direction of bias that the nonresponse of the kind you mentioned might cause (negative, positive or close to zero). No quantification is required, a verbal discussion will suffice.

b) Assume that nonresponse is MCAR (missing completely at random). Estimate the proportion of covid in the area. Estimate also the standard deviation (i.e. square root of the variance) for the estimated proportion.

The area was after the data had been collected divided into subareas depending on risk of contracting covid. People living in close proximity of covid outbreaks had high risk. People not close to outbreaks but in the vicinity of for example bus routes from high risk areas were classified as medium risk. Other areas were low risk. The table below contains data for each subarea.

c) Use data from subareas to estimate the proportion of covid in the whole area. Estimate also the standard deviation (i.e. square root of the variance) for the estimated proportion. It is part of the exam to realise (or deduce) what the numbers in table 2 are and gauge their usefulness. Note that there are small numbers in some subareas. If you believe that there might be a problem with small numbers, state what problem that might be. Also consider some action that may help.

Table 1.

Sub-area	Number of people who live in the sub-area	Number of people who were included in the sample	Number of people who the inter-viewers tested	Number of people who tested positive	Risk
1	1000	50	22	1	1
2	1800	45	22	1	1
3	200	30	15	2	2
4	100	10	5	0	2
5	800	42	30	0	3
6	600	29	20	1	3
7	3500	75	35	6	3
8	2000	50	20	4	3
9	10000	200	100	10	3
whole area	20000	531	269	25	

Table 2.

Sub-area					
1	0.95	0.98	0.043	0.0019	45.5
2	0.98	0.99	0.043	0.0019	81.8
3	0.85	0.93	0.116	0.0071	26.7
4	0.90	0.95	0.000	0.0000	0.0
5	0.95	0.96	0.000	0.0000	0.0
6	0.95	0.97	0.048	0.0023	30.0
7	0.98	0.99	0.142	0.0040	600.0
8	0.98	0.99	0.160	0.0079	400.0
9	0.98	0.99	0.090	0.0009	1000.0
whole area	0.97	0.99	0.084	0.0003	1858.7
Sum of subareas				179349	2183.9

d) Now suppose the statisticians had done the risk classification of subareas *before* designing the SRS and collecting data. What would the best sampling design have been?

Maximum 10 points.

9.

A project team is planning a large survey that will aim at estimating average body mass index in Sweden in the age bracket 15-25. To collect data that would help them to plan the survey, they have conducted a pilot survey with a small simple random sample. The pilot survey sample size was 30. The quantity $s_y^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y}_s)^2$ was found to be 40 000 and $\bar{y}_s$ was 20 in the pilot survey sample.

a) The sample size for the real survey was originally planned to be 1000. With the sample size n = 1000, what would the confidence interval be? Use relevant estimates obtained in the pilot survey.

b) The confidence interval with $n = 1000$ turned out to be too wide. So now the team considers stratified simple random sampling followed by the Horvitz-Thompson estimator within strata. Two equally sized strata are considered. Simplify the estimator $\hat{V}(\bar{y}_{str})$ as far as you can with the information that $N_1 = N_2$. Do not attempt any allocation now. You will not be able to compute a number for $\hat{V}(\bar{y}_{str})$ at this stage; more information is needed.

c) Estimates from the pilot survey are $s_1^2 = 5000$ and $s_2^2 = 20\,000$ in strata 1 and 2. What is n_1 and n_2 with optimal allocation? You will not be able to compute numbers of n_1 and n_2 of with the information given so far, but you can see what k is in $n_1 = kn_2$. Use this to re-express $\hat{V}(\bar{y}_{str})$. (You will still not be able to compute a number for $\hat{V}(\bar{y}_{str})$).

d) For $n = 2000$, what is $\hat{V}(\bar{y}_{str})$? If you have not been able to deduce the value of k in c), use $k = 0.5$.

Maximum 7 points.

Formulas

Red colour refers to Lohr (2022), that is the third edition

Population

Population of size N: $U = \{1, \ \dots, \ i, \ \dots, \ N\}$

Sample, size n: $s = \{1, \ \dots, \ i, \ \dots, \ n\}$

Population total of study variable *y*: $t_y = \sum_{i \in U} y_i$

Population mean of study variable *y*: $\bar{y}_U = \frac{1}{N} \sum_{i \in U} y_i$

Population total of auxiliary variable *x*: $t_x = \sum_{i \in U} x_i$

Population variance: $S_y^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \bar{y}_U)^2$ (Lohr p. 32) 38

A **proportion** is a special case with $y_i = \begin{cases} 1 \text{ if unit } i \text{ has the relevant characteristic} \\ 0 \text{ otherwise} \end{cases}$ (compare Lohr p. 33). 39

For a proportion *P* the population variance $S^2 \approx P(1 - P)$ (Lohr p. 38) 43

Formulas for SRS

Expansion estimator of t_y: $\hat{t}_y = \frac{N}{n} \sum_{i \in s} y_i$

Corresponding estimator of $\bar{y}_U$: $\frac{\hat{t}_y}{N} = \bar{y}_s$

$V(\hat{t}_y) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}$ (Lohr (2.16)) *(2.19)*

For an estimator of $V(\hat{t}_y)$, replace S_y^2 with the following estimator of S_y^2:

$s_y^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y}_s)^2$ (Lohr (2.10) and (2.17) (2.13)(2.20)

Ratio estimator of t_y : $\hat{t}_{rat} = t_x \frac{\hat{t}_y}{\hat{t}_x} = t_x \hat{B}$ (Lohr (4.2)) *(4.3)*

$\hat{V}(\hat{t}_{rat}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n}$, where $s_e^2 = \frac{1}{n-1} \sum_{i \in s} \left(y_i - \hat{B} x_i\right)^2 = s_y^2 + \hat{B}^2 s_x^2 - 2\hat{B} s_{xy}$,

$s_{xy} = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y}_s)(x_i - \bar{x}_s)$ (Lohr (4.8)) (4.10)

It is also ok (even rather better) to use $\hat{V}(\hat{t}_{rat}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n} \left(\frac{\bar{x}_U}{\bar{x}_s}\right)^2$. (Lohr (4.11)) (4.13)

Regression estimator of t_y: $\hat{t}_{reg} = N\left(\bar{y}_s + \hat{B}_1(\bar{x}_U - \bar{x}_s)\right)$, where $\hat{B}_1 = \frac{\sum_{i\in s}(x_i-\bar{x}_s)(y_i-\bar{y}_s)}{\sum_{i\in s}(x_i-\bar{x}_s)^2}$ (Lohr (4.15)) (4.16)

$V\left(\hat{t}_{reg}\right) \approx N^2\left(1-\frac{n}{N}\right)\frac{S_y^2}{n}(1-R^2)$,

where $R = \frac{S_{xy}}{S_x S_y}$ is the finite population correlation coefficient. (Lohr (4.18)) (4.19)

A variance estimator is obtained by replacing the population quantities S_y^2 and R with sample quantities. (Lohr (4.20)) (4.21)

Alternative, equivalent, variance estimator: $\hat{V}\left(\hat{t}_{reg}\right) = N^2\left(1-\frac{n}{N}\right)\frac{s_e^2}{n}$,

where $s_e^2 = \frac{1}{n-1}\sum_{i\in s}\left(y_i - \hat{B}_0 - \hat{B}_1 x_i\right)^2$, $\hat{B}_0 = \bar{y}_s - \hat{B}_1\bar{x}_s$ (Lohr p. 139) 137

Domain estimation in SRS

Let $u_i = y_i x_i$ with $x_i = \begin{cases}1 \text{ if unit } i \text{ belongs to the domain}\\ 0 \text{ otherwise}\end{cases}$ (Lohr p. 134) 140

The part of the sample that falls in domain *d* is denoted by s_d and the number of units in s_d is denoted by n_d.

Estimation of the **mean** of study variable in domain *d*: $\bar{y}_d = \frac{\bar{u}_s}{\bar{x}_s}$

$\hat{V}(\bar{y}_d) = \left(1-\frac{n}{N}\right)\frac{s_{yd}^2}{n_d}$, where $s_{yd}^2 = \frac{\sum_{i\in s_d}(y_i-\bar{y}_d)^2}{n_d-1}$ (compare Lohr (4.13)) (4.22)

Alternatively,

$$\hat{V}(\bar{y}_d) = \left(1-\frac{n_d}{N_d}\right)\frac{s_{yd}^2}{n_d}$$

Estimation of the **total** of study variable in domain *d*, t_d, two cases:

1. If the population size of the domain, N_d, is known: $\hat{t}_d = N_d\bar{y}_d$ (Lohr p. 135) 140
2. N_d is unknown: $\hat{t}_d = N\bar{u}_s$, where $\bar{u}_s = \frac{1}{n}\sum_{i\in s} u_i$. $\hat{V}(\hat{t}_d) = N^2\left(1-\frac{n}{N}\right)\frac{s_u^2}{n}$, where $s_u^2 = \frac{1}{n-1}\sum_{i\in s}(u_i - \bar{u}_s)^2$

Sample size estimation, SRS

We want this precision: $P(|\bar{y}_s - \bar{y}_U| \leq e) = 0.95$. Then, with the approximation $fpc = 1$, $n = \frac{1.96^2 S_y^2}{e^2}$. (compare Lohr (2.25)) (2.31)

Stratification and poststratification

The population is divided into nonoverlapping groups that will exhaust the population fully. I prefer subscript *g* as a generic notation of the number of one poststratum and subscript *h* for a generic notation of the number of one stratum. For example, the sample and total in stratum *h* is denoted by s_h and t_h, respectively. Lohr uses subscript *h* for both kinds of population subsets.

For **stratified simple random sampling** the population mean $\bar{y}_U$ is estimated as

$\bar{y}_{str} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in s_h}\frac{N_h y_i}{n_h} = \frac{1}{N}\sum_{h=1}^{H}\hat{t}_h$ (Lohr (3.1) and (3.2))

and the estimated variance as

$\hat{V}(\bar{y}_{str}) = \sum_{h=1}^{H}\frac{N_h^2}{N^2}\left(1-\frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}$ (Lohr (3.5))

With stratified simple random sampling with proportional allocation, that is, the sample size in each stratum *h* is $n_h = n\frac{N_h}{N}$, the variance of the estimate $\hat{V}(\bar{y}_{str}) = \frac{1}{Nn}\left(1-\frac{n}{N}\right)\sum_{h=1}^{H}N_h s_h^2$ (Lohr p. 86) 90

Optimal allocation, equal costs: $n_h = n\frac{N_h S_h}{\sum_{h=1}^{H}N_h S_h}$ (Lohr (3.14)) (3.15)

For **simple random sampling followed by poststratification**

A variance estimator corresponding to Lohr (3.5) can be used: (3.6)

$\hat{V}(\bar{y}_{post}) = \sum_{g=1}^{G}\frac{N_g^2}{N^2}\left(1-\frac{n_g}{N_g}\right)\frac{s_g^2}{n_g}$

Poststratification estimator of the mean, SRS, general poststratum sample sizes:

$\bar{y}_{post} = \frac{1}{N}\sum_{g=1}^{G}\sum_{i\in s_g}\frac{N_g y_i}{n_g} = \frac{1}{N}\sum_{g=1}^{G}\hat{t}_g$

One-stage cluster sampling, unequal cluster sizes

N and n: number of clusters in the population and in the sample, respectively.

M_i and M_0: number of units in cluster *i* and in the population, respectively.

$t_i = \sum_{j=1}^{M_i}y_{ij}$ is the total of y_{ij} in cluster *i* (y_{ij} is the value of the study variable for unit *j* in cluster *i*).
$\hat{t}_i = t_i$ because in one-stage cluster sampling, all units in the clustered are sampled.

Unbiased estimator of t_y: $\hat{t}_{unb} = \frac{N}{n}\sum_{i\in s}t_i = \frac{N}{n}\sum_{i\in s}\sum_{j=1}^{M_i}y_{ij}$ (Lohr p. 169) 170-171

Corresponding estimator of $\bar{y}_U$: $\hat{\bar{y}} = \frac{\hat{t}_{unb}}{M_0}$ (M_0 must be known)

$\hat{V}(\hat{\bar{y}}) = \frac{N^2}{M_0^2}\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}$, where $s_t^2 = \frac{1}{n-1}\sum_{i\in s}\left(\hat{t}_i - \frac{\hat{t}_{unb}}{N}\right)^2$ (Lohr (5.13)) (5.15)

Ratio estimator of $\bar{y}_U$: $\hat{\bar{y}}_{rat} = \frac{\hat{t}_{unb}}{\hat{M}_0} = \frac{\sum_{i\in s}\hat{t}_i}{\sum_{i\in s}M_i}$

$\hat{V}(\hat{\bar{y}}_{rat}) = \left(1-\frac{n}{N}\right)\frac{1}{n\bar{M}^2}\frac{\sum_{i\in s}(t_i - \hat{\bar{y}}_{rat}M_i)^2}{n-1}$, where $\bar{M} = \frac{1}{n}\sum_{i\in s}M_i$

Horvitz-Thompson estimator

General sampling design, inclusion probability π_i

Unbiased estimator of t_y: $\hat{t}_{HT} = \sum_{i\in s}\frac{y_i}{\pi_i}$ (Lohr (6.19)) (6.25)

Response rate

Response rate computed as $\frac{(6)}{(4)+(3A)}$, where (6) is number of sample units that responds, (4) is number of sample units that are established to be in scope (i.e. belong to target population) and (3A) is the number of unresolved sample units that are believed to be in scope.

Weighting class estimator

$\hat{t}_{WC} = N\sum_{c=1}^{C}\frac{n_c}{n}\bar{y}_{cR}$, where C is number of classes, n_C is sample size in class c, $\bar{y}_{cR} = \frac{\sum_{i\in s_{cR}} y_i}{n_{cR}}$ is the mean of the respondents in class c. (Lohr page 341) 326

Poststratified estimator to adjust for nonresponse

$\hat{t}_{post} = \sum_{g=1}^{G} N_g\bar{y}_{gR}$, where G is number of poststrata, N_g is the population size in poststratum g, $\bar{y}_{gR} = \frac{\sum_{i\in s_{gR}} y_i}{n_{gR}}$ is the mean of the respondents in poststratum g. (Lohr page 342) 331