STOCKHOLMS UNIVERSITET HT2018

Statistiska institutionen
Dan Hedlin

Sample surveys, ST306G

Examination 2018-12-04, 10.00 – 15.00

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.

1.

A self-weighting sample is a sample drawn with a sampling design where all first-order inclusion probabilities are the same.

a) One-stage cluster sampling and simple random sampling are two examples of sampling designs that produces self-weighting samples.
b) With a self-weighting sample, the population mean can be estimated with the unweighted sample mean, if there is no nonresponse.
c) All first-order inclusion probabilities are the same in systematic sampling with random start.
d) The second-order inclusion probabilities in systematic sampling with random start are not the same as the second-order inclusion probabilities in simple random sampling.
e) Because a self-weighting sample is computationally convenient, a sampling design for which all first-order inclusion probabilities are the same is nearly always preferred to other sampling designs.

2.

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.

3.

a) The term overcoverage refers to units that are included in the frame population but not in the target population.
b) The term undercoverage refers to units that belong to the target population but are not included in the frame population.
c) The frame population may contain more units than the actual frame, for example, if the frame is a list of homes for the elderly (addresses) and the target population is the elderly (people at the homes).
d) When computing a nonresponse rate, you have to take undercoverage into account.
e) When computing a nonresponse rate, you are expected to take overcoverage into account.

4.

a) Although quota sampling is a random sampling design, there are potential problems with representativity with this design.
b) Cluster sampling is often used in agricultural crop yield surveys where the purpose is to estimate the total yield of various crops.
c) Two- or multi-stage sampling is often used in social surveys in countries that lack a population register, for example, cluster sampling followed by systematic sampling within sampled clusters.
d) Because Sweden has a very good population register, one-stage sampling designs such as simple random sampling, stratified simple random sampling and systematic sampling are often used in social surveys in Sweden.
e) Systematic sampling may be the best way of drawing a sample of passengers in a train if the purpose is to measure passengers' attitudes towards train services.

5.

a) The intra-class cluster coefficient is a measure of cluster homogeneity.
b) In cluster sampling, the larger the homogeneity within clusters, the lower the variance of estimators.
c) Clusters in cluster sampling can be selected with simple random sampling, systematic sampling or with several other methods.
d) In stratified simple random sampling, one tries, if possible, to make strata homogeneous.
e) If one wants to estimate average progress in a football team through a sample survey, and make separate estimates for players in groups A and B, one should consider stratification by group.

6.

To study the content of calcium in the food offered in homes for the elderly in Washington State, a researcher requested the menus from all the 184 homes for the elderly she could identify in a list of licenced homes in Washington State. The researcher obtained menus from 43 of the 184 homes for the elderly. For simplicity, assume that all homes offered only one choice of dish for each meal and that the researcher selected one day randomly.

a) Give an example of population parameter that you believe is reasonable in this survey.
b) What is the target population, frame population, sampling unit, observation unit and the study variable(s)?
c) Is there any domain of interest (apart from the whole population)?
d) Discuss briefly non-sampling errors in this survey.

Maximum 12 points.

7.

Klein Associates conducted a survey with simple random sampling of 200 solicitors on sexual abuse in the office. 100 solicitors responded, 65% of whom were men and 35% were women. 49% of the women and 9% of the men responded that sexual abuse in the office exists. Ignore nonresponse and the finite population correction. The population size was 2000.

a) Estimate the population difference of the mean of the study variable between women and men and provide a confidence interval for the estimated difference. Ignore the covariance.
b) Klein wants to do survey again, this time using the fact that the frame contains information about the sex (man, woman) of each solicitor. Use the information from their previous survey to compute the minimum sample size to meet the following requirements: the confidence interval should not be wider than 2 times 8 percentage units for women and 2 times 2 percentage unit for men.
c) If you do not ignore the nonresponse, what method could you use in this case to adjust for nonresponse? You need not do any computation, but a one-word answer will not be sufficient.

Maximum 8 points.

8.

A sample of twenty people is drawn in the audience to a political talk. The hall has filled up to its maximum capacity which is an audience size of 300 people. An integer is drawn with equal probability from the numbers 1, 2, 3, …, 300. Denote the number drawn by r. The audience is ordered the following way; people in the front row are ordered from left to right, then the ones in the next row from left to right, and so on until the last row. The people included in the sample are the person who in the ordering process obtained the number r, then person $r + d$, then person $r + 2d$, and so on. If you get to the end of the back row and less than

n people have been included in the sample, where n is the sample size, then you go to the front row and count people from left to right and go on until you have included n people in the sample. In general, for $j = 1,2,\ldots,n$, the sample consists of units $k=r+(j-1)d$ if $r+jd\leq N$, where N is the population size, or $k = r + (j-1)d - N$, if $r+jd>N$. In this survey, $n = 20$ and $d = 15$. Suppose it turned out that $r + 19d = 301$.

a) What is this sampling design called? Be as specific as possible.
b) What is the probability that the person in seat 3 in row 4 is included in the sample?
c) What is the main advantage and main disadvantage with this sampling design compared to the sampling design where all samples with $n = 20$ have the same probability to be selected?
d) Everybody in the sample responds to the survey question (asking about a binary survey variable, "yes" or "no") except the person that was included as the 20th person in the sample. The sample data are 13 people saying yes and 6 saying no. Estimate the proportion of people in the audience who would say "yes", and estimate a confidence interval for that estimate. Make some reasonable assumption about the nonresponse.

Maximum 10 points.

9.

To see if gaming time reduces grades, the head of a large school commissioned a survey. The survey institute selected a simple random sample and reported the following data, where $s_y = \frac{1}{n_r-1}\sum_{i\in s_r}(y_i - \bar{y}_{s_r})^2$ within groups and where s_r and n_r are the set of respondents and number of respondents within groups, respectively. To start with, the head wants an estimate of average gaming time.

Grade	Number of students in the school	Sample size	Number of respondents	Self-assessed number of minutes of gaming	
				$\bar{y}_{s_r}$	s_y
200 or higher	1159	77	61	1.75	0.99
180-199	1163	98	48	4.98	0.91
179 or lower	1678	125	41	9.46	1.53
All	4000	300	150		3.37

a) Ignore the nonresponse and treat the response set of 150 respondents as a simple random sample. Does this procedure correspond to MCAR, MAR or NMAR? Estimate the mean of self-assessed number of minutes of gaming among all student in the school and give a confidence interval for that estimate.
b) Is the assumption about nonresponse in a) realistic? Motivate your answer as specifically as possible.

c) Estimate the mean and a confidence interval with the best nonresponse adjustment method in the present situation. Is your assumption now MCAR, MAR or NMAR?

Maximum 10 points.

Population

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

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

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)

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).

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

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))

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))

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

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

$s_{xy} = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y}_s)(x_i - \bar{x}_s)$ (Lohr (4.8) and (4.11))

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 (4.10) and (4.11))

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))

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

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

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

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. 138-139)

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)

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))

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

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 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)

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

For **simple random sampling followed by poststratification**, if the sample sizes in poststrata are $n_g = n\frac{N_g}{N}$, the variance estimator is the same:

$\hat{V}(\bar{y}_{post}) = \frac{1}{Nn}\left(1 - \frac{n}{N}\right)\sum_{g=1}^{G} N_g s_g^2$ (Lohr (4.22))

For general poststratum sample sizes a variance estimator corresponding to the formula above marked as Lohr (3.5) can be used:

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

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)

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) and p. 170)

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))

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)

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)