STOCKHOLMS UNIVERSITET HT2021

Statistiska institutionen
Dan Hedlin

Sample surveys, ST306G

Examination 2021-11-25, 08.00 – 13.00

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) Suppose an agriculture survey measures the production of rice. If a sampled farm in this survey grows no rice, then this farm is counted as undercoverage.
b) If due to language issues no usable data can be recorded in a telephone interview, then the interview is classified as nonresponse.
c) Suppose it can be established that seven people in a sample of ten people belong to the target population, whereas the status of the other three is uncertain. Six out of the seven respond to the survey. Nobody else responds. If the survey manager believes that only one out the three with uncertain status belongs to the target population, then the nonresponse rate is ¾.
d) Suppose 50 kronor are offered to those of the 1000 sampled people in a survey who have yet not responded and fifty more people respond. Then the nonresponse bias has not necessarily been reduced.
e) Poststratification may reduce nonresponse bias.

2.

a) The decision on what sample size is needed in a survey cannot be based purely on statistical theory; it is also a matter of the purpose of the survey and the required precision of estimates.
b) In general, a survey of a heterogeneous population requires a larger sample size than a survey of a homogeneous population, if everything else is equal.

c) Consider two populations, one with size 500, one with size 500 million. They have the same population variance S_y^2. Then a simple random sample with sample size 50 (10%) from the small population will give better precision for the Horvitz-Thompson estimator of the population mean than a simple random sample with sample size 500 (0.0001%) from the large population, if there is no nonresponse.
d) Suppose pupils in a school survey in an area with 500 schools are asked if they have seen bullying going in their class, with classmates being either perpetrators or victims. For this study variable, a simple random sample of 800 pupils may well give better precision than a one-stage cluster sample of 12 schools with a total sample size of 1000 pupils.
e) To estimate a total with a 95% confidence interval of width $2e$, a sample size of $n = 1.96^2 S_y^2/e^2$ is needed if the *fpc* is ignored (see formulas at the end of this exam paper for notation).

3.

Denote the variance of an estimator by $V(\hat{t})$, which has been called theoretical variance in the course to distinguish it from the variance estimator $\hat{V}(\hat{t})$. Denote the Horvitz-Thompson estimator by $\hat{t}_y$.

a) The theoretical variance of a particular estimator, for example $\hat{t}_y$, follows mathematically from the general definition of variance of an estimator.
b) It is not possible to compute the theoretical variance from a sample.
c) If you want to estimate $V(\hat{t}_y)$, then there are many estimators $\hat{V}(\hat{t}_y)$, with different advantages and disadvantages.
d) A variance estimator may be precise and unbiased.
e) $V(\hat{t})$ is a random variable (stochastic variable).

4.

a) Suppose you want estimates for a subset of a population. Then it is necessary to know before the sample is drawn which units that belong to that subset.
b) The exact choice of size stratum boundaries in stratified simple random sampling is not crucially important if you only want estimates for the whole population.
c) If you want estimates for a domain, then making a stratum that is similar to the domain will make the estimates for the domain more precise.
d) One requirement of a random sample is $\pi_i > 0, \forall i \in U$, that is, positive inclusion probabilities for all units in the population.
e) The requirement of a random sample that $\pi_i > 0, \forall i \in U$, may in practice be deliberately violated.

5.

a) The second-order inclusion probability in simple random sampling is $\frac{n(n-1)}{N(N-1)}$, where *n* is the sample size and *N* is the population size.
b) For some sampling designs the first-order inclusion probability is the same for all units in the population; these sampling designs include simple random sampling and systematic sampling.

c) The first-order inclusion probabilities are the same for all units in the population if the sampling design is stratified simple random sampling with proportional allocation and the stratum sizes are unequal.
d) Let *s* be a sample drawn from a population *U* with simple random sampling and let *r* be the set of *m* respondents. If *r* can be viewed as having been drawn from *s* with simple random sampling, then the probability that a specific unit in *U* is included in *r* is m/N.
e) Suppose the sampling design is one-stage cluster sampling, where clusters are selected with simple random sampling. Then the first-order inclusion probability is $n_I \ / \ N_I$, where n_I is the number of clusters in the sample and N_I is the number of clusters in the population.

6.

Politicians in a certain parliament are offered allowances to cover costs incurred, for example travel expenses. The allowance is paid as a lump sum. As the politicians are not required to declare their expenses, a survey aims to give an estimate on how much of the allowances are actually used to cover necessary expenses. The parliament has $N = 281$ members. The total expenses for each MP (member of parliament) were recorded some years ago. Denote this variable by x_i for MP i. The MPs are the same but their expenses may have changed. A sample of MPs were asked to declare their current expenses. A simple random sample of size 60 was drawn and 50 responded. There are some computed statistics below that you may find useful. The set of 50 responding MPs is denoted by *s*.

a) Due to the nonresponse you need to make some assumption. What assumption? Be as specific as possible.
b) Estimate the mean expenses of the MPs in the parliament using two different estimators, both of which should be either exactly unbiased or approximately unbiased. Estimate also the standard deviation (i.e. square root of the variance) for the estimated mean, for each of the two estimators.
c) Estimate the ratio $\frac{\sum_{i\in U} y_i}{\sum_{i\in U} x_i}$ (no need for variance estimation this time).

$\sum_{i\in U} x_i = 13\ 257; \sum_{i\in s} y_i = 75\ 621; \sum_{i\in s} x_i = 2\ 400; \sum_{i\in s} y_i \ / \ x_i = 1309; \sum_{i\in s} y_i^2 = 270\ 581\ 655; \sum_{i\in s} x_i^2 = 124\ 050; \sum_{i\in s} x_i y_i = 4\ 701\ 177; \sum_{i\in s}\left(y_i - \hat{B}_0 - \hat{B}_1 x_i\right)^2 = 26\ 512\ 463; \sum_{i\in s}\left(y_i - \hat{B}_1 x_i\right)^2 = 950\ 322\ 406; (n-1)S_y^2 = \left(\sum_{i\in s} y_i^2\right) - n\bar{y}_s^2 = 156\ 210\ 942$

Maximum 7 points.

7.

A population *U* consists of four units with labels *a*, *b*, *c* and *d*. The units have the following values of the study variable y.

Unit	*a*	*b*	*c*	*d*
Study variable y	10	10	12	16

A sample of size $n = 3$ is taken. No unit can be drawn twice, that is, a without replacement sampling design is made use of. The sample $\{a, b, c\}$ has probability ½ to be drawn. All other subsets of *U* of size 3 have equal probability of being drawn.

a) Make a list of all possible samples (i.e. those with positive probability of being drawn).
b) What is π_a, that is, the probability that unit *a* is drawn?
c) Which of the following alternatives describes the sampling design best? No motivation is required.
    a. Simple random sampling
    b. Systematic sampling with random start
    c. Cluster sampling
    d. This sampling design has no particular name
d) Suppose the sample is $\{a, b, c\}$. What is the Horvitz-Thompson estimator of the population mean for this sample? If you have not been able to do a) and b), write a formula for the estimator and specify which parts of the formula you do not have values for.
e) Suppose the list of possible samples can be numbered 1, 2, …, *t*, and the Horvitz-Thompson estimates of the population mean are $\bar{y}_{s_1}, \bar{y}_{s_2}, \ldots, \bar{y}_{s_t}$, where for example $\bar{y}_{s_2}$ is the estimated mean based on sample 2 in the list. What is $\sum_{i=1}^{t} \bar{y}_{s_i} \,/\, t$? Note that you do not necessarily need to do any computation to be able to the answer.
f) There are two auxiliary variables as well, see the table below. If you use the auxiliary variable *x*, but not *z*, in a regression estimator, for the sampling design described above, will the variance of this estimator, $V(\hat{t}_{reg})$, be greater or smaller than, or equal to $V(\hat{t}_y)$, where $V(\hat{t}_y)$ is the variance of the Horvitz-Thompson estimator?
g) The other way round, if you use *z*, but not *x*, in a regression estimator, for the sampling design described above, will the variance of this estimator, $V(\hat{t}_{reg})$, be greater or smaller than, or equal to $V(\hat{t}_y)$?

Unit	a	b	c	d
Study variable *y*	10	10	12	16
Auxiliary variable *x*	2	2	2	2
Auxiliary variable *z*	2	2	2	5

Maximum 12 points.

8.

A sample of 500 people is drawn in a Swedish municipality with population size 40 000. There is a population register, which serves as a frame. The variables age and last year's income are known for everyone on the frame. An integer is drawn with equal probability from the numbers 1, 2, 3, …, 80. Denote the number drawn by *r*. The people on the frame are ordered the following way; by age in years, and by income among people of the same age. The people included in the sample are first the person who in the ordering process obtained the number *r*, then person $r + d$, then person $r + 2d$, and so on until 500 people are included in the sample. A questionnaire is sent to the sampled people. The main question, which we

focus on here, reads: “Do you think that the maximum speed of e-scooters should be restricted to 20 km/h?” with response categories yes, no and don’t know.

a) What is this sampling design called?
b) What is a good choice of *d*?
c) What is the probability that Ms Olsson, who is 45 and earned 360 000 kronor last year, is included in the sample?
d) Suppose everybody in the sample responds to the survey question. 40% answered “yes”. Estimate the proportion in the municipality who would answer “yes” and estimate a confidence interval for that proportion. Which assumption(s) did you make to be able to estimate the confidence interval?
e) Suppose now that only 30% of those younger than 35 respond and 40% of those who are 35 or older. Give a formula for and a name of an estimator that you would use to estimate the population proportion of people in favour of a speed limit. Make sure that you have defined all notation. If you find that you cannot compute the estimated proportion, state what further data you would need. State also what of those further data you could obtain from the frame, and what of those further data you could obtain from the respondents. No formulas for variance or confidence interval are needed.

Maximum 13 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. The team computes a confidence interval. As the projected confidence interval is disappointingly wide they consider alternatives.

a) The sample size for the real survey was originally planned to be 1600. With the sample size n = 1600, what would the confidence interval be? Use relevant estimates obtained in the pilot survey.
b) The team wishes to reduce the total width of the confidence interval to 2. They consider increasing the sample to meet that aim. What sample size do they need now? Again, use relevant estimates obtained in the pilot survey.
c) The team considers an auxiliary variable x, known for everybody in the population. This auxiliary variable in a regression estimate is believed to give $R = \frac{S_{xy}}{S_x S_y} = 0.5$. What sample is now needed to make the total width of the confidence interval to 2? Hint: for a Horvitz-Thompson estimator the confidence interval is $\hat{t}_y \pm 1.96\sqrt{\hat{V}(\hat{t}_y)}$ and for a regression estimator the confidence interval is $\hat{t}_{reg} \pm 1.96\sqrt{\hat{V}(\hat{t}_{reg})}$. Use a formula that involves n for $\hat{V}(\hat{t}_y)$ with a formula that involves R for $\hat{V}(\hat{t}_{reg})$.

Maximum 8 points.

Formulae

Population

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

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

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

Horvitz-Thompson 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 $s_e^2 = \frac{1}{n-1} \sum_{i \in s} \left(y_i - \hat{B} x_i\right)^2 = s_y^2 + \hat{B}^2 s_x^2 - 2\hat{B} s_{xy}$,

$s_{xy} = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y}_s)(x_i - \bar{x}_s)$ (Lohr (4.8) 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 $R=\frac{S_{xy}}{S_xS_y}$ 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 $s_e^2=\frac{1}{n-1}\sum_{i\in s}\left(y_i-\hat{B}_0-\hat{B}_1x_i\right)^2$, $\hat{B}_0=\bar{y}_s-\hat{B}_1\bar{x}_s$ (Lohr p. 138-139)

Domain estimation in SRS

Let $u_i=y_ix_i$ 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))

It is also ok to use $\hat{V}(\bar{y}_d)=\left(1-\frac{n_d}{N_d}\right)\frac{s_{yd}^2}{n_d}$.

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)
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 $fpc=1$, $n=\frac{1.96^2S_y^2}{e^2}$. (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

$\bar{y}_{str}=\frac{1}{N}\sum_{h=1}^{H}\sum_{i\in s_h}\frac{N_hy_i}{n_h}=\frac{1}{N}\sum_{h=1}^{H}\hat{t}_h$ (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:

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

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

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

The response rate computed as $\frac{(6)}{(4)+(3A)}$, where (6) is the number of sample units that responds, (4) is the 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*, $\bar{y}_{cR} = \frac{\sum_{i\epsilon s_{cR}} y_i}{n_{cR}}$ 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*, $\bar{y}_{gR} = \frac{\sum_{i\epsilon s_{gR}} y_i}{n_{gR}}$ is the mean of the respondents in poststratum *g*. (Lohr page 342)