STOCKHOLMS UNIVERSITET HT2016

Statistiska institutionen
Jonas Bjermo och Dan Hedlin

Sample surveys, ST306G

Examination 2016-11-29, 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) Stratification is in itself not a sampling design; stratification can be combined with any sampling design within strata.
b) In opinion polls, you often stratify by age, gender and what party the individuals eligible for the opinion poll voted for in the last election.
c) In some surveys, in particular business surveys, the inclusion probability is set to zero in some strata, although that will lead to undercoverage.
d) In some surveys, in particular business surveys, the inclusion probability is set to 1 in some strata, that is, a census is taken in those strata.
e) If you are able to create strata that are internally more homogeneous than the whole population, then the precision of the estimator will be better compared to the precision of an estimator based on a (nonstratified) simple random sample.

2.

a) A systematic sample may be viewed as a cluster sample with only one cluster in the sample.
b) A systematic sample needs a random starting point to be random.
c) A systematic sample may give better precision (that is, smaller variance) than stratified simple random sampling without replacement.
d) A systematic sample may give poorer precision (that is, larger variance) than simple random sampling without replacement, if the values of the study variable show a decreasing trend from the first sampled unit to the last sampled unit.
e) A disadvantage with systematic sampling is the lack of unbiased variance estimators.

3.

a) A domain is a subset of the population.
b) Two domains of interest in one survey cannot be overlapping.
c) Assume that a simple random sampling without replacement is taken from a population register of individuals. If we want to use this sample to estimate the average body mass index of men, then the number of men in the sample is random.
d) Assume that we want to estimate the average body mass index of men. If there is a population register of individuals with their gender recorded, then a stratified simple random sample is preferable to a (nonstratified) simple random sample, because in the former case we can control the number of sampled men.
e) Estimating a domain mean can be viewed as a special case of ratio estimation.

4.

a) If a sample is taken from the Swedish population with the aim to estimate average income, one-stage cluster sampling can be expected to give larger or just as large variance than simple random sampling without replacement.
b) To be efficient (meaning small variance) clusters should, if feasible, be made as heterogeneous in terms of the study variable as possible.
c) Cluster sampling is common in large demographic and health surveys in developing countries.
d) Like strata, clusters divide the population in nonoverlapping subsets that taken together cover the population.
e) If there is a choice and if the costs are equal, it is preferable to have few clusters rather than many clusters.

5.

a) Stratification and poststratification are very common sampling designs.
b) When using weighting classes or poststratification, it is not advisable to construct to small classes or poststrata. They should contain at least about 20 observation units.
c) One reason to use poststratification is reduced nonresponse bias, which will be the result of poststratification if there are auxiliary variables that are associated with the probability of responding.
d) A continuous variable can be used as auxiliary variable in poststratification if it is first categorised into groups; however, without any transformation a continuous variable cannot be used as auxiliary variable in poststratification.
e) Like stratification, poststratification divides the population into nonoverlapping subsets that taken together cover the population; however, one difference is that stratification must be done before the sample is taken, poststratification can be done after the sample has been taken and data collected.

6.

A simple random sample without replacement of size 1500 was taken from the part of the Swedish automobile register that contains cars smaller than 1500 kg. The car owners, according to the register, were matched with the Swedish population register, and the number

of children of the car owners was extracted from the population register. The table below reports on number of owners by number of children in two groups: owners to cars classified as environmentally friendly and owners to other cars. (Maximum 8 points.)

	Number of owners	
Number of children	**Environmentally friendly cars**	**Other cars**
0	76	528
1	139	189
2	166	225
3	63	76
4	19	8
5+	9	2
Sum	472	1028

a) Is the sample of owners a simple random sample? If not, why?
b) Assume from now on that the sampled owners constitute a simple random sample. Estimate the proportion of owners of environmentally friendly cars who have children. Estimate also the variance of the estimate of the proportion. Do the same for "other cars".
c) Estimate the difference of the proportions in b) and the variance of the difference. If you need to make an approximation when estimating the variance, make sure you state it explicitly.

7.

A municipality decided that they wanted to estimate the grade of dementia among their elderly living in retirement homes. Three among ten retirement homes were randomly chosen and trained doctors examined all the elderly living in the chosen retirement homes. According to developed procedures every elderly's grade of dementia were put on a scale from 0 to 100. The elderlies' test results are given in the table below. (Maximum 10 points.)

	Retirement home		
Patient	**1**	**2**	**3**
1	30	10	30
2	20	50	27
3	96	50	10
4	20	47	70
5	30	60	26
6	63	30	83
7	60	67	45
8	20	25	
9	55	30	
10	71	7	
11	53		
12	45		
13	20		
14	5		
Sum	588	376	291

a) What is this sampling design called?
b) Three types of nonsampling error are undercoverage, nonresponse and measurement error. For each, give one likely reason why this sampling error may occur in this survey, or, if there is no likely reason why a specific type on nonsampling error should occur, explain why.
c) Estimate the mean test result in the population, using the estimator that gives the smallest variance, and give the standard error of the estimator.

8.

A frame of size 1000 is stratified into two strata. Each stratum contains 500 individuals. A stratified simple random sample of size 4 from each stratum is drawn. There are two frame variables, called B and x. In stratum 1, 400 individuals have value B=1 and the remaining 100 individuals have value B=2. In stratum 2, 300 individuals have value B=1 and 200 individuals have value B=2. The total of x is 600 in stratum 1 and 700 in stratum 2. The sample data are in the table below. Unfortunately one value of y was not recorded due to no contact despite twelve contact attempts. Seven individuals, whose values of y were recorded, were verified to belong to the target population. (Maximum 10 points.)

a) Estimate the mean of the study variable y by using 1) the weighting class estimator, 2) poststratification, 3) hot deck imputation and, finally, as method 4) ignore the record with the missing value. The sample mean of y is 3.5 in stratum 1 and 3 in stratum 2 if the record with the missing value is ignored. Note that there are some choices to be made in these calculations.
b) What is the response rate? Give an answer in the form of an interval, smallest to highest.

Sample from strata 1 and 2

Stratum	B	x	y
1	1	1	1
1	1	2	3
1	2	1	4
1	2	1	6
2	1	1	1
2	1	3	.
2	2	1	3
2	2	1	5

9.

In a city there are 25 petrol stations. A frame of employees was compiled and a simple random sample without replacement was taken from each of three categories of employees. From the sample data we obtained the table below. (Maximum 12 points.)

a) Estimate the mean salary of station managers in this population and estimate standard error (standard deviation) for this estimate.
b) Which method of allocation has been applied?

c) The survey institute wants to repeat the survey, but this time they believe that they can draw a better sample with the information that they now have. Suggest sample sizes that should give the best precision. The total sample size should still be 32.

d) The survey institute wants also an estimate of the total sample size that would give an error (difference between estimated mean salary and true mean salary) which would be less than 4000 with high probability. Compute the sample size they will need. You may ignore the finite population correction.
You may find the following formula helpful:
$\sum_{i=1}^{N}(y_i - \bar{y}_U)^2 = \sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_i - \bar{y}_{U_h})^2 + \sum_{h=1}^{H} N_h(\bar{y}_{U_h} - \bar{y}_U)^2$, where $\bar{y}_U$ is the population mean, and $\bar{y}_{U_h}$ and N_h are the stratum mean and stratum size. Further, the sample sum of squares within strata is 479 298 455.

e) [Try to answer this even if you have not answered d)] The finite population correction may be a fair bit smaller than 1. What is the implication of ignoring the finite population correction in the sample size computation?

Category	Number of employees at 25 petrol stations	Sample size	Mean salary in the sample	Variance of the salaries in the sample, $s_h^2 = \frac{1}{n_h-1}\sum_{i=1}^{n_h}(y_i - y_{s_h})^2$	s_h
Station managers	35	7	36 029	25 553 025	5055
Sales staff	85	17	14 865	3 888 784	1972
Other staff	40	8	20 088	29 289 744	5412
Sum	160	32			

Formulae

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

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

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: $P(|\bar{y}_s - \bar{y}_U| \leq e) = 0.95$. 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, 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 $\bar{y}_U$: $\hat{\bar{y}}_{rat} = \frac{\hat{t}_{unb}}{M_0} = \frac{\sum_{i\in s}\hat{t}_i}{\sum_{i\in s}M_i}$

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