STOCKHOLMS UNIVERSITET HT2017

Statistiska institutionen
Dan Hedlin

Sample surveys, ST306G

Examination 2017-11-23, 16.00 – 21.00

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) 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) 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.
d) The first-order inclusion probabilities are not 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.
e) In one-stage cluster sampling, 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.

2.

a) One meaning of the term representative (as in 'a representative sample') in surveys might be that the sample is a miniature of the population. However, many good sampling designs that are widely used do not produce samples that are representative in this sense of the term.
b) In design-based inference (which we have focused on in the course), the inclusion probabilities play a crucial role in the inference.

c) In surveys using non-random sampling designs an issue is often the lack of knowledge of the inclusion probabilities; however, if all first-order and second order inclusion probabilities are known, non-random samples allow for valid inference.
d) The Horvitz-Thompson estimator of a population total is unbiased.
e) For simple random sampling, the variance of the Horvitz-Thompson estimator of a population total is $V(\hat{t}_y) = N^2\left(1-\frac{n}{N}\right)\frac{S_y^2}{n}$, where n is the sample size, N is the population size and S_y^2 is the population variance.

3.

a) Eurostat's framework for quality of survey statistics (which we have focused on in the course) contains several components, three of which are goodness-of-fit, accuracy and comparability.
b) Accuracy consists of two parts: sampling errors and nonsampling errors.
c) Badly phrased questions lead to measurement error, which is a type of nonsampling error.
d) Nonresponse may, or may not, lead to nonresponse bias, which is a type of nonsampling error.
e) Undercoverage is the result of some units in the target population having zero inclusion probability.

4.

a) 'Domain estimation' refers to the estimation of areas or spaces.
b) The exact choice of size stratum boundaries in stratified simple random sampling is not very important if you only want estimates for the whole population.
c) If you want estimates for domains, then strata that are similar to the domains will make the estimates for those domains more accurate.
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 $\pi_i > 0, \forall i \in U$, that is, positive inclusion probabilities for all units in the population, is in practice often deliberately violated.

5.

a) Cluster sampling is often used in school surveys.
b) 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.
c) 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.
d) Systematic sampling may be the best way of drawing a sample of passengers in a train.
e) Quota sampling is a random sampling design commonly used in opinion polls.

Maximum 10 points.

6.

A union is interested in how much mining businesses spend on staff training. They obtain a reliable frame of all active mining businesses. They select a simple random sample from the 42 mining businesses on the frame and manage to obtain responses from all twelve businesses in the sample. The union wishes to estimate the ratio of the mean cost of training per full-time employee, that is, their parameter of interest is t_y/t_x, where t_y is the population total of cost for staff training and t_x is the number of full-time employees in the mining industry.

The sample sums are 627.7 for variable y and 93.2 for variable x. Other sums are

$\sum_{k=1}^{12} \frac{y_k}{x_k} = 20.2$; $\sum_{k=1}^{12} x_k^2 = 994.78$; $\sum_{k=1}^{12} y_k^2 = 51\,498.09$; $\sum_{k=1}^{12} x_k y_k = 6\,917.59$.

The population variance is $S_y^2 = \frac{1}{N-1}\sum_{i\in U}(y_i - \bar{y}_U)^2 = \frac{1}{N-1}\left[\sum_{i\in U} y_i^2 - N\bar{y}_U^2\right]$.

A useful variance formula is $V(\hat{B}) = \left(1 - \frac{n}{N}\right)\frac{s_e^2}{n\bar{x}^2}$ where $\bar{x}$ is the sample mean of variable x. For other notation, see appropriate formulae at the end of this exam paper.

a) Estimate the parameter of interest and give a 95% confidence interval for the estimate.
b) A statistician tells the people at the union that they can make good use of the variable x, which registered on the frame for all businesses. The union decides to use the fact that t_x=320. Use this fact to compute an unbiased estimate of t_y/t_x and give a 95% confidence interval for this estimate.
c) Compare the confidence intervals in parts a) and b). Which one is wider? Is this surprising? (Even if you have not been able to compute both intervals, you can still guess, with motivation, which of them should be wider.)
d) Suggest a better sampling design. Note that you can do this even if you have not been able to compute estimates in parts a) and b).

Maximum 10 points.

7.

We have drawn a simple random sample of 500 people from a population with 12 000 people. One variable is total bank savings. The sum of the bank savings for the people in the sample was 50 million euro. The population standard deviation was estimated as 1 million euro.

a) Estimate the total bank savings in the population and a 95% confidence interval.
b) How large should the sample size have been for the total length of the confidence interval around the estimated total to be at most 1 000 million? Assume 100% response.
c) After having published the results of the survey, we learn that another institute did the same survey at the same time, but with a simple random sample from each of two subpopulations: 1) people registered as being unemployed and 2) everybody else. 8.33% of the population have registered as being unemployed. Estimate the total bank savings in the population and a 95% confidence interval. The sample data were:

Subpopulation	Sample size	Sum bank saving for those in sample (million euro)	Population standard deviation, estimate (million euro)
Unemployed	100	10	0.5
Others	400	88	0.8

d) Compare the length of this confidence interval with the one you computed in part a). Why is it shorter (or wider, if it is wider)? (Even if you have not been able to compute both intervals, you can still guess, with motivation, which of them should be wider.)

Maximum 10 points.

8.

A sample of twenty students is drawn among the students sitting in a lecture hall. It is known that 300 students are there. An integer is drawn with equal probability from the numbers 1, 2, 3, …, 300. Denote the number drawn by r. The students are ordered the following way; the students 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 students included in the sample are the student who in the ordering process obtained the number r, then student $r + d$, then student $r + 2d$, and so on until either $r + 19d \leq 300$ or $r + 19d > 300$, where $n = 20$ and $d = 15$. Suppose it turned out that $r + 19d = 301$. Then the student to the far left in the front row is included in the sample.

a) What is this sampling design called? Be as specific as possible.
b) In row three there is a guy with an angry face. What is the probability that this guy is included in the sample?
c) Everybody in the sample responds to the survey question, except the guy with the angry face, who actually was included in the sample but refused to take part. The question is: 'Did you go to the university today by bicycle?" (response categories yes or no). The sample data are 13 students saying yes and 6 saying no. Estimate the proportion of students in the lecture hall who went to the university by bicycle, and estimate a confidence interval for that estimate. Make some reasonable assumption about the nonresponse. Why do you think your assumption is reasonable?

Maximum 10 points.

9.

Children in 46 schools were immunised against measles. Parents had to give their consent before an injection could be given. They were given a form where they could agree to immunisation against measles. However, many parents did not return the consent form. We

want to know what proportion of the non-immunised children did not obtain the vaccination because their parents did not return the consent form.

A sample of ten schools was selected with simple random sampling and each sampled school counted the number of non-immunised children and how many of those had parents who had not returned the form.

a) What is this sampling design called?
b) Using some (or all) of the data given on next page, estimate the relevant target parameter. Estimate also the variance of the estimate.
c) Assume that the data you based your computation on in part b) had been obtained with a simple random sampling design. So consider a simple random sample of children that has been drawn from all children in the 46 schools. Estimate the variance of the proportion estimate, if the data are analysed as a simple random sample. (Hint. First: what formula are you going to use? Second: make a list of the quantities in the formula and what values they have in this case. Third: compute the variance.)
d) Compare the variances in parts b) and c). Which one is larger? Is this surprising? (Even if you have not been able to compute both variances, you can still guess, with motivation, which of the variances should be larger.)

Maximum 10 points.

School	Number of children	Number of non-immunised children	Number of non-immunised children whose parents did not return the consent form	The ratio of column 2 to column 4	The ratio of column 3 to column 4
1	181	78	37	4.89	2.11
2	645	238	119	5.42	2.00
3	732	261	179	4.09	1.46
4	401	174	104	3.86	1.67
5	498	236	94	5.30	2.51
6	502	188	98	5.12	1.92
7	131	113	74	1.77	1.53
8	200	170	83	2.41	2.05
9	302	296	179	1.69	1.65
10	214	207	69	3.10	3.00
Sum	**3806**	**1961**	**1036**	3.67	1.89

In the table below, y_i is the number non-immunised children whose parents did not return the consent form, z_i is the number of children in the school, x_i is the number of non-immunised children in the school, $r = \hat{t}_y / \hat{t}_z$, $p = \hat{t}_y / \hat{t}_x$, where $\hat{t}_y$ is the estimated total of y.

School	$y_i - rz_i$	$(y_i - rz_i)^2$	$y_i - px_i$	$(y_i - px_i)^2$
1	-12.268	150.52	-4.208	17.70
2	-56.570	3200.18	-6.736	45.37
3	-20.251	410.13	41.113	1690.30
4	-5.153	26.55	12.076	145.82
5	-41.556	1726.94	-30.680	941.22
6	-38.645	1493.46	-1.321	1.74
7	38.341	1470.08	14.302	204.54
8	28.559	815.65	-6.811	46.39
9	96.795	9369.28	22.623	511.78
10	10.749	115.54	-40.358	1628.81
Sum	**0.0000**	**18778.33**	**0.0000**	**5233.68**

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

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

$s_{xy} = \frac{1}{n-1}\sum_{i \in s}(y_i - \bar{y}_s)(x_i - \bar{x}_s)$ (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: $\hat{V}(\hat{t}_{reg}) = N^2\left(1-\frac{n}{N}\right)\frac{s_e^2}{n}$,

where $s_e^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \hat{B}_0 - \hat{B}_1 x_i)^2$, $\hat{B}_0 = \bar{y}_s - \hat{B}_1\bar{x}_s$ (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, $\bar{y}_{cR} = \frac{\sum_{i \epsilon s_{cR}} y_i}{n_{cR}}$ 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)