STOCKHOLMS UNIVERSITET HT2021

Statistiska institutionen
Dan Hedlin

Sample surveys, ST306G

Examination 2021-10-29, 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) MAR is a natural assumption when using simple random sampling and the Horvitz-Thompson estimator, if the data collection suffers from nonresponse.

b) Nonresponse is a serious and an increasingly worrying issue in the Labour force survey (arbetskraftsundersökningen) conducted by Statistics Sweden.

c) The most realistic assumption in social surveys that suffer from nonresponse is probably not missing at random (NMAR).

d) The reason why MAR is more often made use of than NMAR, is that NMAR makes estimation complex, difficult or impossible.

e) Missing completely at random, MCAR, is a stronger assumption (that is, an assumption less likely to be satisfied in practice) than missing at random, MAR.

2.

a) When you use the formula for optimal allocation in stratified simple random sampling, it is not uncommon that the formula results in $n_h > N_h$ for a stratum; that is, the sample size is larger than the population size in that stratum according to the optimal allocation formula.

b) One disadvantage with the systematic sampling design is that the variance cannot be estimated without bias.

c) If you take a sample of five passengers in a bus (in all, there are $N > 5$ passengers in the bus), the set of samples that can be drawn with systematic sampling is a subset of the set of samples that can be drawn with simple random sampling of passengers.

d) The inclusion probabilities in stratified simple random sampling followed by poststratification may be equal or unequal within strata.

e) Choices to make when designing a stratified sample include

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

3.

a) One reason to use stratified sampling is to be protected from the possibility of obtaining a really bad sample.

b) One reason to use stratified sampling is to make sure of obtaining a good sample size in an important subgroup of the population.

c) One benefit of using simple random sampling in a national survey on income is that you cannot obtain a sample with mostly high-income earners.

d) Clusters in cluster sampling are subsets of the population.

e) Domains are subpopulations that may or may not overlap.

4.

a) Blind pursuit of high response rates is unwise.

b) Collecting auxiliary data (one or several auxiliary variables) on respondents and nonrespondents is the key to useful nonresponse adjustments.

c) Poststratification and regression estimation may reduce nonresponse bias.

d) If using imputation, it is wise to apply several methods for imputation to see if they produce very different results or not.

e) It is not possible to use the same auxiliary variable in both sampling design and estimation.

5.

a) One disadvantage with register-based statistics compared with a census, is that the register may not contain exactly the variables you need.

b) In general a census is more expensive than a sample survey, which is more expensive than producing the same statistics with register-based statistics (provided that register-based statistics is feasible).

c) It is not possible to use more than one auxiliary variable (e.g. age) in the regression estimator, if the aim is to estimate mean income.

d) A regression estimator is, in principle, a Horvitz-Thompson estimator plus an adjustment.

e) Even if the correlation between the study variable and the auxiliary variable is poor, the regression estimator is likely to be more precise (yield smaller variance) than the Horvitz-Thompson estimator.

6.

There is a need to estimate the total yield of runner beans (and other crops) in country X. There is a frame of farms that contains an indicator variable, which takes the value 1 for farms that grew runner beans five years ago and zero for other farms. The frame also contains the yield of runner beans six years ago. However, this variable is not fully reliable since some farms have value 1 on the indicator and still zero yield. The National office for agriculture consider two alternative data collection methods: 1) mail out a questionnaire to a simple random sample of farms 2) send out people to a simple random sample of farms. At the farm they would divide the beds where runner beans are grown into plots and take a random sample of two plots. The runner beans in the sampled plots would be harvested and weighed.

a) What is the target population? Be as specific as possible.
b) What is the primary sampling unit for data collection method 2?
c) What is the observation unit for data collection method 1 and the observation unit for data collection method 2? (Another term for observations unit is reporting unit.)
d) Which of the data collection methods do you think is best for this survey? Ignore costs. Assume that the same number of farms are sampled in both methods. Consider precision and potential sources of bias.
e) Discuss briefly what estimators that could be made use of if data are collected with method 1, and if you suggest more than one, which one you would prefer.

Maximum 12 points.

7.

The Swedish Statistical Society and its two sections want estimates of the proportion and number of their members who are female. Here we assume that every member belongs to exactly one section (in reality many members belong to two sections or no section) and that there are only two sections, the Survey Association and the Cramér Society. The Survey Association took a simple random sample of their members and so did the Cramér Society. Assume that everybody responded. (Fictitious numbers)

	Membership	Sample size	Female members in sample
The Survey Association	2000	200	80
The Cramér Society	500	200	40
Sum	2500	400	120

a) What is this sampling design called?
b) Using the data above, estimate the proportion and the number of women in the Statistical Society and in each section, that is, in the Survey Association and in the Cramér Society. Estimate also the variance for the estimated proportion in the Statistical Society.

c) Is it reasonable to say that the Survey association has a greater proportion of women than has the Cramér society? You do not need to make a formal hypothesis test, but you do need to support your conclusion with appropriate estimates.

d) The Survey Association wants an estimate of mean income for their new members. Three of the sampled members are new; their incomes are 28 000, 35 000 and 36 000. Estimate the mean income for new members and estimate also the variance for the estimated mean. You may (or may not) find that the values of these quantities are useful: $\sum_{i\in s_d}(y_i - \bar{y}_d)^2 = 20000$ and $\sum_{i\in s}(y_i - \bar{y}_s)^2 = 350000$.

Maximum 12 points.

8.

A travel agency intends to ask a simple random sample of customers a binary question: interested or not interested in a particular kind of special offer. They want a sample large enough to be able to produce a 95% confidence interval with the total width 4 percentage units. They believe that 30% of the members in the target population may be interested in the offer. The customer database, which will be used as the frame, has 100 000 records. The database holds one record for each person who has ever provided the agency with her or his name, an email address and a phone number.

a) How large should the sample be?

b) If they have no idea of what proportion of the customers that may be interested, what sample size should the agency take to be on the safe side?

c) Having drawn a simple random sample of 100 records from the database, and obtained 50 responses, the agency finds that 30% of the respondents are indeed interested in the special offer. Estimate the proportion of all customers that would be interested in the offer. State what assumption(s) you make, if any.

d) Discuss briefly potential reasons for the fact that the agency obtained a response from only 50%.

Maximum 10 points.

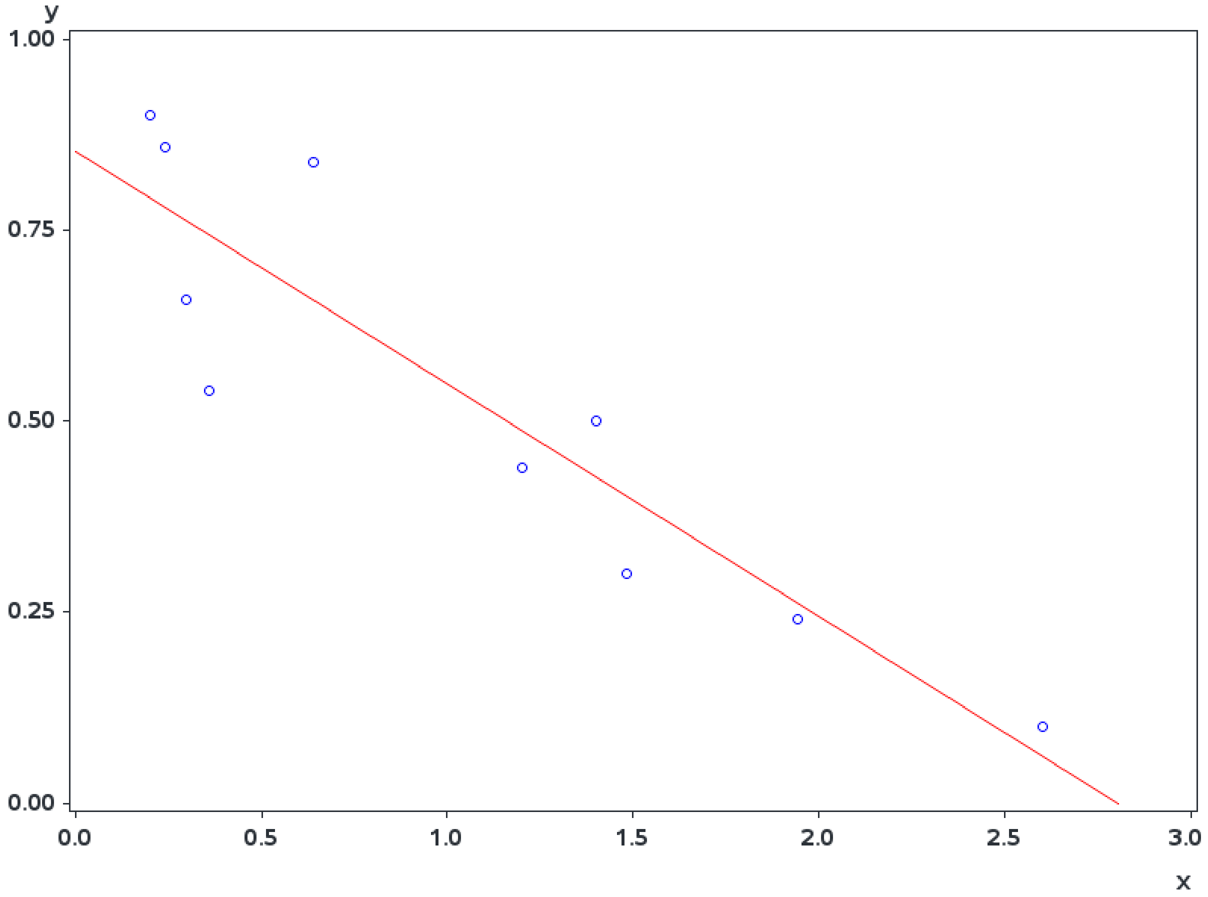
9.

A simple random sample of size 10 is drawn from a population of size 1000. The aim is to estimate the total $t_y = \sum_{k=1}^{1000} y_k$. The sample data are:

y_k	0.10	0.24	0.30	0.44	0.50	0.54	0.66	0.84	0.86	0.90
x_k	0.60	0.31	0.02	0.33	0.44	0.95	0.30	0.52	0.40	0.60

a) Compute two estimates of t_y using two different estimators, both of which should be either exactly unbiased or approximately unbiased. Estimate also the variances of the two estimates. You may find some of these evaluated sums useful: $\sum_{k=1}^{10}(y_k - \bar{y})(x_k - \bar{x}) = 0.13$, $\sum_{k=1}^{10}(x_k - \bar{x})^2 = 0.54$, $\sum_{k=1}^{10}(y_k - \bar{y})^2 = 0.69$, $\sum_{k=1}^{10} x_k = 4.47$, $\sum_{k=1}^{10} y_k = 5.38$, $\sum_{k=1}^{1000} x_k = 447$, $\frac{s_{xy}}{s_x s_y} = 0.216$, $s_y^2 = 0.08$, $s_x^2 = 0.06$

b) Now suppose the sample data are as in the graph below. With these data, how would you rank the estimators you used in a), and why? You do not necessarily need to compute anything, apart from what you did in a).



Maximum 6 points.

Formulae

Dan Hedlin, Department of Statistics, Stockholm University

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

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

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

and the variance as

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

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:

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

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

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)