STOCKHOLMS UNIVERSITET HT2019

Statistiska institutionen
Dan Hedlin

Sample surveys, ST306G

Examination 2019-10-29, 15.00 – 20.00

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) Systematic sampling is a sampling design in which every *kth* unit (e.g. every tenth) on the frame is included in the sample.

b) In cluster sampling the frame is partitioned into non-overlapping groups and the sample consists of a sample of those groups ('partition' means that every unit on the frame belongs to one, and only one, group).

c) If a systematic sample is selected from a frame, the order of the objects in the frame is important.

d) When there is a choice between different sampling designs in a survey, the main issue is what frame the statistician has access to, because once the characteristics of the frame is given (frame units and variables), then the sampling design is also given.

e) Systematic sampling is actually a form of cluster sampling where only one cluster is selected.

2.

a) The second-order inclusion probability in simple random sampling without replacement is $\frac{N}{n} \cdot \frac{N-1}{n-1}$, 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 second-order inclusion probabilities are important, because they appear in the general formula for estimation of the variance of the mean, and hence also in the formulae for specific sampling designs.

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 the type of nonresponse is missing completely at random, MCAR, then r can be viewed as having been drawn from s with simple random sampling, which implies that the probability that a specific unit in U is included in r is m/N.

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. That is, all secondary sampling units have the same first-order inclusion probability.

3.

a) Sometimes the variance of the Horvitz-Thompson estimator can be larger, and even far larger, for stratified simple random sampling with proportional allocation than for simple random sampling without stratification. This typically happens if the strata are not carefully defined.

b) To use stratification in one-stage sampling, the population size for each stratum, N_h, and the stratum membership for every unit on the frame must be known.

c) Design issues in stratification include defining strata, choosing the total sample size and allocating the total sample to the strata.

d) In optimal allocation, the aim is to minimise the variance of the estimator for a given total cost (or the other way round: to minimise cost for given variance).

e) Disproportional allocation refers to allocation of sampling units to strata so that the sampling fractions n_h / N_h are not equal for all strata.

4.

a) If a survey contains sensitive questions, the estimates may suffer from measurement errors.

b) If the sample is large and you are not interested in domain estimation, bias resulting from non-sampling errors is usually a bigger problem than large variance.

c) One reason to use mixed modes (for example, a mixture of telephone interviews, web questionnaires and postal paper-and-pen questionnaires) is to increase the response rate.

d) If a labour force survey targets people between 15 and 74 years of age, and if you interview someone and realise that this person is actually 78 but has been included in the sample due to an error in the register, then the error is classified either as a measurement error or nonresponse.

e) If someone responds to a survey question about what party she or he voted for in the latest election, and the respondent by mistake mentions a party that she or he did not vote for, that error is classified as a measurement error.

5.

a) Common population parameters in official statistics are proportions, means and totals.

b) You often see statements like 'the wealthiest 1% of the population own more than the poorest 75%'. If you define the total wealth of the wealthiest 1% as $t = \sum_{i \in U} y_i u_i$, where u_i=1 if individual i belongs to the wealthiest 1% of the population and u_i=0 otherwise, then t is a population parameter (or a domain parameter).

c) If you use mixed modes (for example, a mixture of telephone interviews and postal paper-and-pen questionnaires) to estimate the proportion of unemployed, then the proportion of unemployed among those who have been interviewed by telephone is a population parameter important to assess nonresponse bias.

d) Suppose the people in the population register, between 15 and 74 years of age, define the target population in a labour force survey. If you estimate both employment in general and employment among people who have come onto the register in the first half-year of 2019, then 'people who have entered the population register between 1 January 2019 and 31 July 2019' is a domain.

e) Domains are subsets of the target population, and to facilitate the estimation of domain parameters, stratification is often made use of.

6.

The aim of a survey that a yeast factory conducts is to estimate the amount of a substance that by law must be below a certain threshold (toxic in large doses). A systematic sample of size 10 is drawn from a batch of packages of yeast. The batch contains 1000 packages. The aim is to estimate the mean $\bar{y}_U = \frac{1}{1000}\sum_{k=1}^{1000} y_k$ where y_k is the amount of the substance in micrograms. There is another substance in yeast, which often shows positive correlation with the substance of interest. Denote the amount of this second substance by x_k. The sample data are:

y_k	0.1	0.24	0.3	0.44	0.5	0.54	0.66	0.84	0.86	0.90
x_k	0.24	0.64	0.3	0.76	0.36	1.4	1.2	1.48	1.94	2.60
k	1	2	3	4	5	6	7	8	9	10

a) Use three different estimators to compute three estimates of $\bar{y}_U$. All estimators should be either exactly unbiased or approximately unbiased. Estimate also the variances of the three estimates. For the variance estimation, use SRS variance formulas (as is common for systematic samples). You may find some of these evaluated quantities useful:
$\frac{1}{9}\sum_{k=1}^{10}(y_k - \bar{y})(x_k - \bar{x}) = 0.196, \frac{1}{9}\sum_{k=1}^{10}(x_k - \bar{x})^2 = 0.606, \frac{1}{9}\sum_{k=1}^{10}(y_k - \bar{y})^2 = 0.090,$
$\sum_{k=1}^{10}(y_k - \bar{y})^2 = 0.812, \ \sum_{k=1}^{10} x_k = 10.92, \sum_{k=1}^{10} y_k = 5.41,$
$(\sum_{k=1}^{10} y_k)^2/(\sum_{k=1}^{10} x_k)^2 = 0.245, \sum_{k=1}^{1000} x_k = 1000.$

b) Someone in middle management notes that four of the sample values are higher than the others. He back-tracks those packages and finds that a two of them, numbers 7 and 9, were produced between 10.31 and 10.47 am, and the other two, numbers 8 and 10, between 11.43 and 11.57 am. He wants a domain estimate of the mean of all packages produced during those two periods of time. The number of packages produced during those two periods of time is not known. Compute that domain estimate, without using the amounts of the second substance denoted by x_k. No variance estimate is needed.

c) When the middle manager obtains the domain estimate, he believes that it is too high. Although it is below the legal threshold, which is 100 micrograms, it suggests unsafe variability, he thinks. He gives the order that the batch of 1000 packages be destroyed. Critique the statistical rationale of his strategy.

Maximum 10 points.

7.

There is a population consisting of the four units in the table below.

Id number	y
1	10
2	10
3	12
4	16
Total	48

With the sampling design that has been decided, four samples are possible to draw: {1,2,3}, {1,2,4}, {1,3,4}, and {2,3,4}, each with probability 1/4. Now suppose that one of the possible samples has been drawn. We want to estimate the mean of y. As you can see, the population mean is 12. **Note:** not all of the items a-h below need computation; if you answer any of them without computation, make sure that you motivate your answer.

What is (see formulae at the end of the exam for notation)

a) the inclusion probability of the unit with id number 3

b) the population sum of the inclusion probabilities, $\sum_{i \in U} \pi_i$

c) the expected value of the Horvitz-Thompson estimator $\hat{t}_{HT}$

d) the variance of $\hat{t}_{HT}$

e) the bias of $\hat{t}_{HT}$

f) the mean squared error of $\hat{t}_{HT}$

g) What does $\sum_{i \in s} 1/\pi_i$ estimate?

h) Does this sampling design have a specific name; if so, what is this sampling design called?

Maximum 10 points.

8.

A sample of ten greenhouses was selected with simple random sampling and in each sampled greenhouse the number of tomato plants affected by an insect that hampers the growth of the tomatoes was counted. The sample was taken from some geographically adjacent farms with 46 greenhouses in total. The aim was to estimate the proportion of plants affected by the insect.

a) What is this sampling design called?

b) Use some or all of the data given on the next page to estimate the relevant population parameter. Estimate also the variance of the estimate.

c) Suggest a sampling design that may have given lower variance and would have been fairly easy to implement. Make sure that you motivate your suggestion well.

d) Suppose the first value of the number affected plants, 37, is missing. What would be the best way to deal with that? You do not need to do any computation.

Maximum 10 points.

Greenhouse	Number of plants in greenhouse	Number of plants affected by the insect	The ratio of column 2 to column 3
1	82	37	0.451
2	238	119	0.500
3	261	179	0.684
4	170	104	0.612
5	236	94	0.400
6	188	98	0.521
7	113	74	0.655
8	170	83	0.489
9	296	179	0.605
10	207	69	0.333
Sum	**1961**	**1036**	0.528

In the table below, y_i is the number of plants affected by the insect, z_i is the number of all tomato plants in the greenhouse, $r = \hat{t}_y \,/\, \hat{t}_z$, where $\hat{t}_y$ and $\hat{t}_z$ are the estimated totals of y and z. The square of the mean number of plants in the ten greenhouses is 38 455.

Greenhouse	y_i-rz_i	$(y_i - rz_i)^2$
1	-6.3208	39.95
2	-6.7358	45.37
3	41.1132	1690.30
4	14.1887	201.32
5	-30.6792	941.22
6	-1.3208	1.74
7	14.3019	204.54
8	-6.8113	46.39
9	22.6226	511.78
10	-40.3585	1628.81
Sum	**0.0**	**5311.43**

9.

A simple random sample without replacement is drawn from a population of size $N = 400$. The sample size is $n = 40$. The values of two study variables, denoted by y and z, are observed for all sample units. The aim of the survey is to estimate the difference between the estimated totals. It was found that $\sum_{i \in s} y_i = 900$, $\sum_{j \in s} z_j = 400$, $S_y^2 = 42$, $S_z^2 = 29$. Below, there are is a formula and a computed statistics that may be useful.

a) Estimate the totals t_y and t_z and the variances of their estimates, $\hat{t}_y$ and $\hat{t}_z$.

b) Estimate the difference, which can be estimated as simply as $\hat{\Delta} = \hat{t}_y - \hat{t}_z$, and then estimate the variance of $\hat{\Delta}$.

c) Interpret $\hat{\Delta}$ in the light of $\hat{V}(\hat{\Delta})$. If you have not done b), assume that $\hat{V}(\hat{\Delta}) = 200\,000$.

d) Suppose that one sample is drawn from population U_1 and another sample is drawn from another population U_2. Both samples are drawn with simple random sample without replacement, and both have size $n = 40$. The variable y is observed in the first sample, and the variable z in the second one. What is $\widehat{V}(\widehat{\Delta})$ now? If you need to make a reasonable assumption, make sure you state it.

$\text{Cov}(\hat{t}_y,\hat{t}_z)=N^2\left(1-\frac{n}{N}\right)\frac{S_{yz}^2}{n}$, where $S_{yz}^2=\frac{1}{N-1}\sum_{i\in U}\sum_{j\in U}(y_i-\bar{y}_U)(z_j-\bar{z}_U)$. $S_{yz}^2 = 1.3$

Maximum 10 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

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

$\bar{y}_{str} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in s_h}\frac{N_h y_i}{n_h} = \frac{1}{N}\sum_{h=1}^{H}\hat{t}_h$ (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 $\bar{y}_U$: $\hat{\bar{y}}_{rat} = \frac{\hat{t}_{unb}}{\hat{M}_0} = \frac{\sum_{i\in s}\hat{t}_i}{\sum_{i\in s}M_i}$

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

Department of Statistics

Stockholms universitet

Correction sheet

Date: 29/10 - 2019

Room: Värtasalen

Exam: Sample Surveys

Course: Sample Surveys

Anonymous code: 0001-PXK

☒ I authorise the anonymous posting of my exam, in whole or in part, on the department homepage as a sample student answer.

NOTE! ALSO WRITE ON THE BACK OF THE ANSWER SHEET

Mark answered questions

1	2	3	4	5	6	7	8	9	Total number of pages
✗	✗	✗	✗	✗	✗	✗	✗	✗	5
Teacher's notes				8	9	9	10	10	

1k

Points	Grade	Teacher's sign.
46	A	

SU, STATISTIK

Skrivsal: Värtasalen Anonymkod: 0001-PXK Blad nr: 1

1. d) R

2. a) R

3. a) R

4. d) R

5. ~~d)~~

8

SU, STATISTIK

Skrivsal: Värtasalen **Anonymkod:** 0001-PKK **Blad nr:** 2

6. a) HT-estimator

$$\hat{\bar{Y}}_{HT} = \frac{\hat{t}_y}{N} = \frac{541}{1000} = \boxed{0,541}$$ R

$$\hat{t}_y = \frac{N}{n}\sum_{i\in s} y_i = \frac{1000}{10}\cdot 5,41 = 541$$

$\sum_{i\in s} y_i = 5,41$ ← from the question

$$\hat{V}(\hat{\bar{Y}}_{HT}) = \left(1-\frac{n}{N}\right)\frac{S_y^2}{n} = \left(1-\frac{10}{1000}\right)\frac{0,09}{10} = \boxed{0,00891}$$ R

$$S_y^2 = \frac{\sum_{i=1}^{10}(y_i-\bar{y})^2}{n-1} = 0,090$$ ← from the question

Ratio-estimator

$$\hat{\bar{Y}}_{rat} = \bar{X}_U \cdot \frac{\bar{y}_s}{\bar{x}_s} = \left(\frac{1000}{1000}\right)\cdot\frac{\left(\frac{5,41}{10}\right)}{\left(\frac{10,92}{10}\right)} \approx \boxed{0,495}$$ R

choosing this variance formula →

$$\hat{V}(\hat{\bar{Y}}_{rat}) = \left(1-\frac{n}{N}\right)\frac{S_e^2}{n} = \left(1-\frac{10}{1000}\right)\frac{0,0444}{10} \approx \boxed{0,00440}$$ R

$$S_e^2 = S_y^2 + \hat{B}^2 S_x^2 - 2\hat{B}S_{xy} = 0,090 + 0,495^2\cdot 0,606 - 2\cdot 0,495\cdot 0,196 \approx 0,0444$$

from the question →

$$S_{xy} = \frac{1}{9}\sum_{i=1}^{10}(y_i-\bar{y})(x_i-\bar{x}) = 0,196$$

$\hat{B} \approx 0,495$ from above since $\bar{X}_U = 1$

$$S_x^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i-\bar{x})^2 = 0,606$$

$$S_y^2 = \frac{1}{9}\sum_{i=1}^{10}(y_i-\bar{y})^2 = 0,090$$

Regression-estimator

$\frac{10,92}{10}$

$$\hat{\bar{Y}}_{reg} = \bar{y}_s + \hat{B}_1(\bar{x}_U - \bar{x}_s) = 0,541 + 0,323(1-1,092) \approx \boxed{0,511}$$ R

$$\hat{B}_1 = \frac{\sum_{i\in s}(x_i-\bar{x}_s)(y_i-\bar{y}_s)}{\sum_{i\in s}(x_i-\bar{x}_s)^2} = \frac{\cancel{9}\cdot 0,196}{\cancel{9}\cdot 0,606} \approx 0,323$$

$$\hat{V}(\hat{\bar{Y}}_{reg}) \approx \left(1-\frac{n}{N}\right)\frac{S_y^2}{n}(1-r^2) = \left(1-\frac{10}{1000}\right)\frac{0,090}{10}(1-0,839^2) \approx \boxed{0,00263}$$ R

6

$$r = \frac{S_{xy}}{S_x S_y} = \frac{0,196}{\sqrt{0,606}\sqrt{0,090}} \approx 0,839$$

b) and c) →

Two individual domains and one combined domain.

b) Produced 10.31 - 10.47 am N_d unknown

$$\hat{Y}_{d_1} = \frac{\sum_{i \in S_{d_1}} Y_i}{n_{d_1}} = \frac{0{,}66+0{,}86}{2} = 0{,}76$$

← extra (misunderstood the question at first)

Produced 11.43 - 11.57 am

$$\hat{Y}_{d_2} = \frac{\sum_{i \in S_{d_2}} Y_i}{n_{d_2}} = \frac{0{,}84+0{,}90}{2} = 0{,}87$$

Combined 10,31-10,47 am and 11,43-11,57 am

$$\hat{Y}_{d_3} = \frac{\sum_{i \in S_{d_3}} Y_i}{n_{d_3}} = \frac{0{,}66+0{,}84+0{,}86+0{,}9}{4} = 0{,}815$$

← this estimate is wanted R

1

c) First, one can discuss what is "unsafe variability", a vague term in my opinion. Also, he s selected a small sample, not only for the total (n=10), but also for the domains he is interested in (individually $n_{d_i} = 2$, $i=\{1,2\}$, combined $n_{d_3} = 4$). This makes it difficult to say anything about the population since we get wide confidence intervals (if calculated). Thus, it is a strange statistical rational he has, not only because the sample estimate is below the threshhold, but also due to what I wrote in the beginning of this paragraph

2

9

SU, STATISTIK

Skrivsal: Vårtasalen Anonymkod: 0001-PKK Blad nr: 3

$\pi_i = \pi = \frac{n}{N}$ konstant

$\frac{1}{\pi} = \frac{N}{n}$

7. $\bar{y}_U = 12$

Sample	Probability	Z_3	Z_1	Z_2	Z_4	$\sum_s y_i$	$\frac{N}{n}\sum_s y_i = \hat{t}_y$
1 2 3	1/4	1	1	1	0	32	42,67
1 2 4	1/4	0	1	1	1	36	48
1 3 4	1/4	1	1	0	1	38	50,67
2 3 4	1/4	1	0	1	1	38	50,67

$Z_3 = \begin{cases} 1, & \text{if unit 3 is included in sample} \\ 0, & \text{otherwise} \end{cases}$

a) $\pi_3 = \frac{1}{4}\cdot 1 + \frac{1}{4}\cdot 0 + \frac{1}{4}\cdot 1 + \frac{1}{4}\cdot 1 = \frac{3}{4}$ R 1

The inclusion probability of unit 3 is $\frac{3}{4}$.

b) $\sum_{i\in U} \pi_i = \frac{3}{4} + \frac{3}{4} + \frac{3}{4} + \frac{3}{4} = \frac{12}{4} = 3$ R 1

Since all units have the same inclusion probability, since they are included in 3 samples each, with probability 1/4 => $\pi_i = 3 \cdot 1/4 = 3/4$

The population sum of the inclusion probabilities is 3.

c) $E(\hat{t}_{HT}) = \sum P(s)\,\hat{t}_y = \frac{42,67}{4} + \frac{48}{4} + \frac{2\cdot 50,67}{4} = 48$ R 1

d) $\hat{V}(\hat{t}_{HT}) = \sum P(s)(\hat{t}_y - t_y)^2 = \frac{1}{4}(42,67-48)^2 + \frac{1}{4}(48-48)^2 + \frac{2}{4}(50,67-48)^2 =$

$t_y = 48$ $\approx 10,67$ R 1

e) $\text{Bias} = E(\hat{t}_{HT}) - t_y = 48 - 48 = 0$ unbiased R 1

f) $MSE(\hat{t}_{HT}) = \hat{V}(\hat{t}_{HT}) + (\text{Bias})^2 = 10,67 + 0^2 = 10,67$ R 1

g) $\frac{1}{\pi_i} = w_i = \frac{N}{n}$

R 2

$\sum_{i\in s} \frac{1}{\pi_i}$ measures the sum of how many units in the population, including the units in the sample, one sampled unit is "representing", i.e. the sum of the sampling weights.

h) No, it has no name. No partition, since belonging to more groups, ie not cluster sampling or stratified SRS. Not systematic. Closest would be SRS for selecting the sample as a whole. It is SRS 1

9

SU, STATISTIK

Skrivsal: Vartasalen Anonymkod: 0001-PXK Blad nr: 4

8. $n=10$ SRS $N=46$

a) It is called one-stage cluster sampling, using SRS to select clusters. R 2

b) $\hat{p}_{rat} = \frac{\hat{t}_{unb}}{\hat{M}_0} = \frac{\sum_{i \in S} t_i}{\sum_{i \in S} M_i} = \frac{1036}{1961} \approx 0,5283$ R

$$\hat{V}(\hat{p}_{rat}) = \left(1-\frac{n}{N}\right)\frac{1}{n\bar{M}^2} \cdot \frac{\sum_{i \in S}(t_i - \hat{p}_{rat}M_i)^2}{n-1} = \left(1-\frac{10}{46}\right)\frac{1}{10 \cdot 196{,}1^2} \cdot \frac{5311{,}43}{9} \approx$$

$\approx 0{,}0012$ R 2

$\bar{M} = \frac{1}{n}\sum_{i \in S} M_i = \frac{1}{10} \cdot 1961 = 196{,}1$

r in table is $\hat{p}_{rat}$.

c) One alternative is stratified simple random sampling, dividing the farms into strata in terms of areas. Though, they are geographically adjacent (from question), there might still be possibilities to divide the farms into smaller stratas, inorder to get a lower variance.

However, as there are few farms and we are interested in plants, a better idea would maybe be to draw a systematic sample of plants. There is no need for a list since you can see for yourself when entering a greenhouse how many plants there are. The reason why this would reduce the variance is because all farms is included in the sample with some plants. Also as the plants are in haphazard order, this won't be a problem. The disadvantage is that you might not know how large sample should be if there is no list of how many plant each farm has.

Although, these two ideas for sampling plants has some disadvantages, they might be easy to implement and reduce the variance. If I had to choose I would use stratified SRS, if possible, to select farms and then use either systematic or clustering to select plants. 3

d) Using regression imputation to compute an estimation of affected plants by insect a farm with 82 plants →

would have. This is done by using the information from all other sampled clusters.

3

10

SU, STATISTIK

Skrivsal: Värtasalen Anonymkod: 0001-P~~X~~K Blad nr: 5

9. $N=400$ SRS $n=40$

a) $\hat{t}_y = \frac{N}{n}\sum_{i\in S} y_i = \frac{400}{40}\cdot 900 = 9000$

$\hat{V}(\hat{t}_y) = N^2\left(1-\frac{n}{N}\right)\frac{s_y^2}{n} = 400^2\left(1-\frac{40}{400}\right)\frac{42}{40} = 151200$ R

$\hat{t}_z = \frac{N}{n}\sum_{j\in S} z_j = \frac{400}{40}\cdot 400 = 4000$

$\hat{V}(\hat{t}_z) = N^2\left(1-\frac{n}{N}\right)\frac{s_z^2}{n} = 400^2\left(1-\frac{40}{400}\right)\frac{29}{40} = 104400$ R

2

b) $\hat{\Delta} = \hat{t}_y - \hat{t}_z = 9000 - 4000 = 5000$ R

$\hat{V}(\hat{\Delta}) = \hat{V}(\hat{t}_y - \hat{t}_z) = \hat{V}(\hat{t}_y) + \hat{V}(\hat{t}_z) - 2\cdot \mathrm{Cov}(\hat{t}_y, \hat{t}_z) =$

$= 151200 + 104400 - 2\cdot\left(N^2\left(1-\frac{n}{N}\right)\frac{s_{yz}^2}{n}\right) =$

$= 151200 + 104400 - 2\cdot\left(400^2\left(1-\frac{40}{400}\right)\cdot\frac{1,3}{40}\right) =$

$= 246240$ R

c) $\hat{\Delta}$ is the estimated difference of t_y and t_z using $\hat{t}_y$ and $\hat{t}_z$ as estimates of t_y and t_z. One way to interpret $\hat{\Delta}$ in the light of $\hat{V}(\hat{\Delta})$ is by the coefficient of variation: $\hat{cve} = \frac{\sqrt{246240}}{5000} \approx 0,0992$, which is a way to see the spread in the estimate. It is not unreasonably high but one could wish for an even lower $\hat{cve}$, the lower the $\hat{cve}$, the better. One can also calculate a confidence interval (95%): $5000 \pm 1,96\sqrt{246240}$

$\hat{cve} = \frac{\hat{se}}{\hat{\theta}}$

0 is not included in C.I. i.e. statistically significant difference

$[4027 ; 5973]$ Ok

Using these two ways to interpret $\hat{\Delta}$ in the light of the variance, we can see that our estimate is not superprecise, however the precision is not super low. It is somewhere inbetween, a decent estimate.

d) In this case I assume that the samples have been drawn independently, resulting in me assuming the covariance between y and z to be 0. This gives:

$\hat{V}(\hat{\Delta}) = \hat{V}(\hat{t}_y - \hat{t}_z) = \hat{V}(\hat{t}_y) + \hat{V}(\hat{t}_z) - 0 = 151200 + 104400 = 255600$

Note: I also assume the population sizes to be 400 still in both populations. Ok

10