STOCKHOLMS UNIVERSITET

Statistiska institutionen
Dan Hedlin

HT2018

Sample surveys, ST306G

Examination 2018-10-30, 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) A domain is a subpopulation for which estimates are desired.
b) If the sampling frame contains information that identifies (exactly or approximately) important domains, then that information can be used in stratification to create strata that are similar to the domains.
c) In b), if two domains overlap, then one can let two strata overlap, if they are similar to the domains.
d) Two choices to make when designing a stratified sample is what sampling design or sampling designs to use and how to allocate the sample to strata.
e) In poststratification, the sample is divided into subgroups based on characteristics of the sample units.

2.

a) If you draw a sample of time slots (e.g. hours) in order to interview all visitors to an institution during the sampled slots, the calendar of opening hours can be viewed as a frame.
b) The sampling design in a) is a cluster sample.
c) If the purpose is to obtain estimates for all visitors during one calendar month, and you choose not to sample time slots in one weekend in the targeted month (zero inclusion probability for those slots), the resulting statistical error is called overcoverage.

d) Suppose the data from one of the interviewers, who has interviewed ten visitors, don't make sense. If you decide to discard the data from this interviewer, the resulting statistical error is called nonresponse (or possibly measurement error).
e) If you in d), however, decide to edit and use the edited data from the interviewer who don't seem to have done a good job, then the statistical error associated with these data is called measurement error.

3.

a) A census is a survey in which the aim is to measure the entire population.
b) Register-based statistics use administrative registers to produce statistics.
c) If you mail questionnaires to all members of the Riksdag (the parliament) to find out who they believe should be the next prime minister, you are in fact conducting a census.
d) A census may suffer from nonsampling errors.
e) Suppose the aim is to conduct a survey of members of the Riksdag (the parliament). If the intention is to mail a questionnaire to all members, but by mistake the list of names does not include all members of the Riksdag, then the resulting statistical error is called sampling error.

4.

a) Auxiliary variables are used in estimation to reduce sampling and nonsampling errors.
b) Regression estimation and poststratification are two common ways of using auxiliary variables in estimation.
c) Poststratification is also a sampling design method.
d) The regression estimator may reduce nonresponse bias.
e) If there is a choice between the ratio and the regression estimator, the regression estimator is usually preferable.

5.

a) If there are no cost-reasons or administrative or practical reasons against, it is better to design a cluster sample with few clusters than many clusters.
b) In cluster sampling, the primary sampling units are the unit that are sampled from the frame.
c) If the sole criterion is to minimise variance, the clusters in cluster sampling should be as heterogeneous as possible.
d) If the sole criterion is to minimise variance, the strata in stratified simple random sampling should be as homogeneous as possible.
e) Cluster sampling rarely gives smaller variance than simple random sampling.

6.

In a survey on how many times people have been victims of a crime during the reference period, a simple random sample of 900 individuals was selected from the population register of a Swedish town. The size of the population eligible for the survey was 63 000. 400 responded. Use the assumption MCAR.

Number of times an individual has been victim of crime	Number of respondents	Number of women	Number of men
0	290	170	120
1	75	30	45
2+	25	10	15
Sum	390	210	180

a) Is MCAR realistic or unrealistic? Only a brief motivation is required.
b) Using the data above, estimate the proportion of women who have been victimised and give the standard error for the estimated proportion. You may use reasonable approximations, but make sure you state them
c) Is it fair to say that the data in the table above suggest that women are victims of crime more often than men? Assume that the covariance is zero.
d) Would you say that the data ***below*** show that men are victims of crime more often than women? Assume that the covariance is zero. You do not need to make a formal hypothesis test.

Number of times an individual has been victim of crime	Number of respondents	Number of women	Number of men
0	500	280	240
1	75	30	25
2+	25	10	15
Sum	600	320	280

Maximum 10 points.

7.

There are two alternative plans for a survey with the same total budget, denoted by C_T. Let C_U be the fixed cost of the survey. The total cost for the survey is $C_T = C_U + C_R n_R + C_{NR} n_{NR}$, where C_R is the cost per respondent, n_R is the number of respondents, C_{NR} is the cost per sample unit that does not respond and n_{NR} is the number of nonrespondents. The total sample size is $n = n_{NR} +$

n_R. The sampling design will be simple random sampling. The target parameter is the mean of a continuous variable y. The population size is larger than 1 000 000. The total cost C_T must not exceed 1000 euro. Focus on sampling error and nonresponse bias. The aim is to minimise the MSE (mean squared error) for a fixed budget.

Here are the alternative plans:

1. Inexpensive follow-up of initial nonrespondents. Unconditional incentive to everybody. $C_T = 1000$, $C_U = 100$, $C_{NR} = 1.5$, $C_R = 6$. The variance of the estimated mean is expected to be 100, the proportion $\frac{n_R}{n}$ is expected to be $\frac{1}{9} \approx 0.11$ and the bias is expected to be in the interval 10-12.
2. No incentive and intensive follow-up. $C_T = 1000$, $C_U = 100$, $C_{NR} = 5$, $C_R = 5$. The variance of the estimated mean is again expected to be 100, the proportion $\frac{n_R}{n}$ is expected to be $\frac{1}{3} \approx 0.33$ and the bias is expected to be in the interval 1-2.

a) For each alternative, compute the smallest and largest values that the MSE can take and decide which of the alternatives is preferable. (If you fail to compute the MSE, try to decide with some other argument which alterative is preferable.)
b) For each alternative, compute the sample size. Hint: express the right-hand-side of $C_T = C_U + C_R n_R + C_{NR} n_{NR}$ as a function of n by observing that $n_R = n\left(\frac{n_R}{n}\right)$

Maximum 6 points.

8.

A food inspector samples packages of Swedish meat delivered to one Ica Maxi store to estimate the total and the mean of worm fragments in the meat. In total there are 580 boxes. She selects 12 boxes with simple random sampling without replacement and inspects every package in the 12 boxes. The data and some statistics are given below.

a) What is this sampling design called?
b) However, the count for box 5 was lost. Apply two reasonable methods to address this issue. The inspector suspects that box 5 might be similar to box 4 because they come from the same farm. On the other hand, she argues with herself, this farm shouldn't be any different than any other meat producing farm in this delivery.
c) Using the data below, estimate the total of worm fragments in the 580 boxes and provide 95% confidence interval for the total. Do this for both methods you chose to address the missing value.
d) Do you expect that your variance estimates suffer from negative bias, positive bias or that they are unbiased? Only a brief motivation is required.

Maximum 12 points.

Box	1	2	3	4	5	6	7	8	9	10	11	12
Number of worm fragments	13	10	3	15	-	10	11	5	15	8	20	0

$$\Sigma(y_i-\bar{y})^2=\Sigma y_i^2-\frac{(\Sigma y_i)^2}{n}$$

$$\left(\left(\sum^{n-1} y_i\right)+y_j\right)^2=\left(\sum^{n-1} y_i\right)^2+y_j^2+2y_j\sum^{n-1} y_i$$

$$\sum_{i=1}^{11} y_i=110 \qquad \sum_{i=1}^{11} y_i^2=1438 \qquad \left(\sum_{i=1}^{11} y_i\right)^2=12100 \qquad 1-\frac{11}{580}\approx 0.98 \qquad 1-\frac{12}{580}\approx 0.98$$

9.

A simple random sample of size 31 is taken from a forest with 2967 pine trees. The sample data and some sample statistics are given below. The sum of all diameters of pines in the forest is 1062 meters.

a) What is the estimated ratio of the volume to the diameter?

Estimate the total volume of the pine trees, in cubic meters, and the square root of the variance of that estimate using

b) the Horvitz-Thompson estimator
c) the ratio estimator
d) the regression estimator.
e) Judging from the graph below, explain the ordering of the variances of the regression estimator, the ratio estimator and the Horvitz-Thomson estimator.

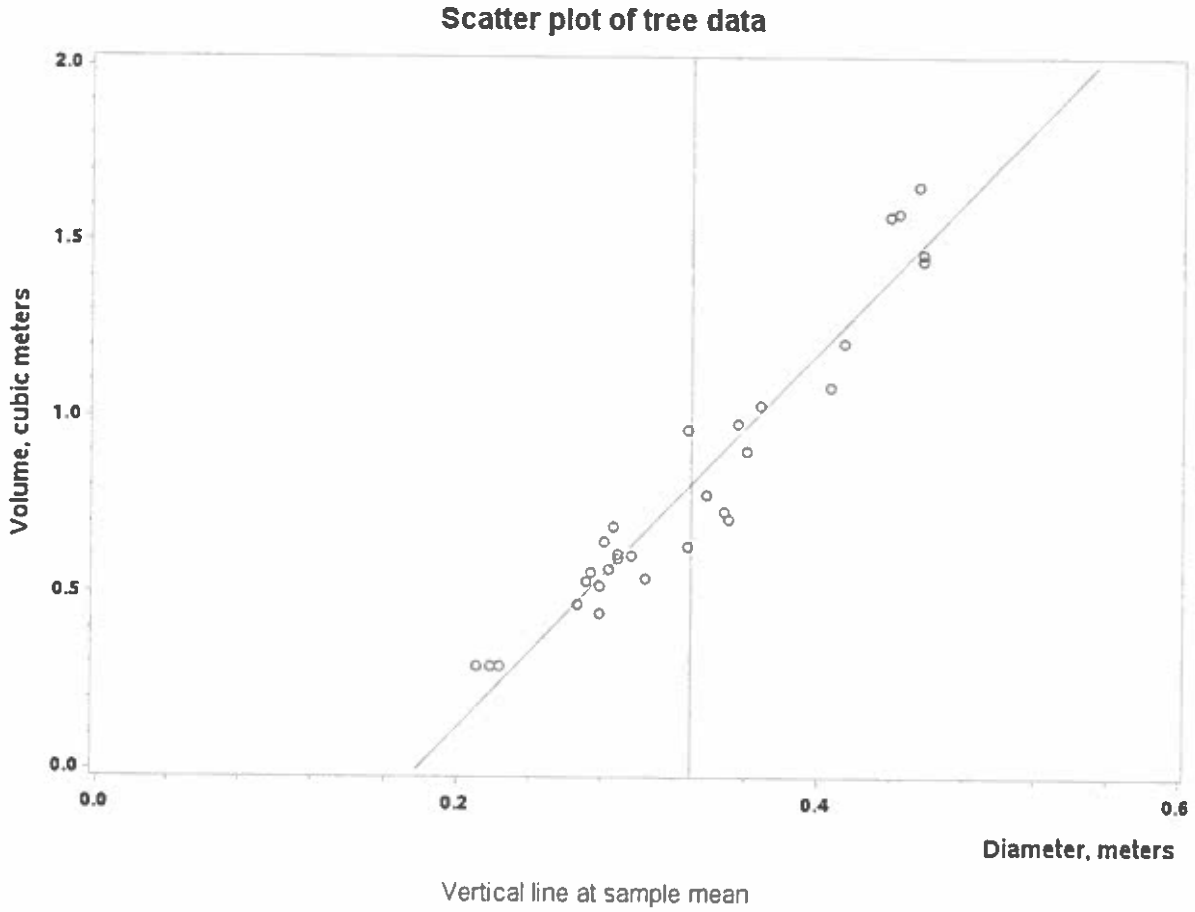
Maximum 12 points.

Sample sum of diameters, m	10.4
Sample sum of volumes, m^3	26.5
An estimate of the finite population correlation coefficient	0.94
s^2 of the diameters in the sample	0.006
s of the diameters in the sample	0.08
s^2 of the volumes in the sample	0.22
s of the volumes in the sample	0.47

$\hat{B}_1 = \frac{\sum_{i \in S}(x_i - \bar{x}_S)(y_i - \bar{y}_S)}{\sum_{i \in S}(x_i - \bar{x}_S)^2}$ where x_i is the diameter of pine i and y_i is the volume of pine i	5.66

Label	diameter	volume
1	0.2108	0.2922
2	0.2184	0.2922
3	0.2235	0.2894
4	0.2667	0.4653
5	0.2718	0.5334
6	0.2743	0.5589
7	0.2794	0.4426
8	0.2794	0.5164
9	0.2819	0.6412
10	0.2845	0.5646
11	0.2870	0.6866
12	0.2896	0.5958
13	0.2896	0.6072
14	0.2972	0.6043
15	0.3048	0.5419
16	0.3277	0.6299
17	0.3277	0.9590
18	0.3378	0.7774
19	0.3480	0.7292
20	0.3505	0.7065
21	0.3556	0.9789
22	0.3607	0.8994
23	0.3683	1.0299
24	0.4064	1.0867
25	0.4140	1.2087
26	0.4394	1.5718
27	0.4445	1.5804
28	0.4547	1.6541

Label	diameter	volume
29	0.4572	1.4612
30	0.4572	1.4470
31	0.5232	2.1847
Sum	**10.4318**	**26.5369**



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

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

$V(\hat{t}_{reg}) \approx N^2\left(1-\frac{n}{N}\right)\frac{S_y^2}{n}(1-R^2)$,

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*: $\bar{y}_d = \frac{\bar{u}_s}{\hat{x}_s}$

$\hat{V}(\bar{y}_d) = \left(1 - \frac{n}{N}\right)\frac{s_{yd}^2}{n_d}$, 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 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: 30/10/2018

Room: Brunnsvikssalen

Course: Sample surveys (eng)

Exam: Sample surveys (eng)

Anonymous code: 0012-CDS

☒ 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
X	X	X	X	X	X	X	X	X	3
Teacher's notes 10					10	6	11	8	

16

Points	Grade	Teacher's sign.
45	A	

SU, DEPARTMENT OF STATISTICS

Room: BR Anonymous code: 0012-CDS Sheet number: 01

(1-5)

1 C

2 C

3 E

4 C

5 A R

10

(6)

a MCAR is not realistic, more than half of the sample did not respond. For the assumption missing completely at random to hold true $Cov(\pi, \phi)$, where π is inclusion probability and ϕ is response probability was be 0. It is unrealistic that the fact that you have been victim of a crime does not affect your want to answer to a survey about it. 24

b The simple proportion: $\hat{p} = \frac{\sum y_i}{n} = \frac{30+10}{210} = \underline{0,1905}$

Estimated population variance $\hat{V}(\hat{p}) = \left(1 - \frac{n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1} =$

$= \left(1 - \frac{210}{63000/2}\right) \frac{0,1905(1-0,1905)}{210-1} \approx 0,0007329$ $SE = \underline{0,02707}$ 24

We assume 50% women in population

c Estimated proportion difference: $\hat{d}_{mw} = \hat{p}_m - \hat{p}_w$

$\hat{p}_m = \frac{60}{180} \approx 0,\overline{3}$ $\hat{d}_{mw} = 0,\overline{3} - 0,1905 \approx 0,142833$

$\hat{V}(\hat{d}_{mw}) = \hat{V}(\hat{p}_m - \hat{p}_w) = \hat{V}(\hat{p}_m) + \hat{V}(\hat{p}_w) - \underbrace{2Cov(\hat{p}_m, \hat{p}_w)}_{=0} = 0,0007329 + 0,001234 \approx 0,001967$

$\hat{V}(\hat{p}_m) = \left(1 - \frac{180}{63000/2}\right) \frac{\frac{1}{3}\left(1-\frac{1}{3}\right)}{180-1} \approx 0,001234$

(c) cont.

$H_0: \hat{d}_{mw} < 0$ Test statistic: $z_{obs} = \frac{0,14283}{\sqrt{0,00467}} \approx 3,220$

$H_1: \hat{d}_{mw} \geq 0$ Critical value: $z_{crit} = 1,96 \Rightarrow$ Reject H_0

Significance 5%

We reject H_0, the data does not suggest that women are victims of crime more often than men. 1V

N.b. "More often" is interpreted as proportion of the population, not the times each gender has been victimized 0V

d. $\hat{d}^*_{mw} = \hat{p}^*_m - \hat{p}^*_w = \frac{40}{290} - \frac{40}{320} \approx 0,01766$

$\hat{V}(\hat{d}^*_{mw}) = \left(1 - \frac{290}{31500}\right) \frac{\frac{40}{290}\left(1 - \frac{40}{290}\right)}{274} + \left(1 - \frac{320}{31500}\right) \frac{\frac{40}{320}\left(1 - \frac{40}{320}\right)}{314} \approx 0,0007743$

$SE(\hat{d}^*_{mw}) \approx 0,0278$

The difference is less than one standard error from 0, the data does not show that men are victimized more often. 0V

10

SU, DEPARTMENT OF STATISTICS

Room: BR Anonymous code: 0012-CDS Sheet number: 02

(7)

C_T = total budget C_R = cost/respondent C_{NR} = cost/non respondent

C_U = fixed cost n_R = respondent count n_{NR} = non respondent count

$C_T = C_U + C_R n_R + C_{NR} n_{NR}$ $n = n_R + n_{NR}$

$C_T \leq 1000$

a) Alt 1. $C_T = 1000$ $C_U = 100$ $C_{NR} = 1,5$ $C_R = 6$ $\frac{n_R}{n} = \frac{1}{9} \Rightarrow \frac{n_{NR}}{n} = \frac{8}{9}$

$1000 = 100 + 6 \cdot n_R + 1,5 \cdot 8 \cdot n_R$

$900 = 18 \cdot n_R \Rightarrow n_R = 50$ $n_{NR} = 8 n_R$

$SSE_{Bias} = s_1^2 \cdot (n_R - 1) = 4900$

Bias 10:

$SSE_{Bias10} = 4900 + 10^2 \cdot 50 = 9900$

$MSE_{Bias10} = \frac{9900}{49} \approx 202$

Bias 12:

$SSE_{Bias12} = 4900 + 12^2 \cdot 50 = 12100$

$MSE_{Bias12} = \frac{12100}{49} \approx 247$

Alt 2. $C_T = 1000$ $C_U = 100$ $C_{NR} = 5$ $C_R = 5$ $\frac{n_R}{n} = \frac{1}{3} \Rightarrow \frac{n_{NR}}{n} = \frac{2}{3}$

$1000 = 100 + 5 n_R + 5 \cdot 2 n_R$

$900 = 15 n_R \Rightarrow n_R = 60$ $n_{NR} = 2 n_R$

$SSE_{Bias} = s_2^2 \cdot (n_R - 1) = 5900$

Bias 1:

$SSE_{Bias1} = 5900 + 1^2 \cdot 60 = 5960$

$MSE_{Bias1} = \frac{5960}{59} \approx 101$

Bias 2:

$SSE_{Bias2} = 5900 + 2^2 \cdot 60 = 6140$

$MSE_{Bias2} = \frac{6140}{59} \approx 104$

Alt 1 MSE can be between 202 and 247 about right

Alt 2 MSE can be between 101 and 104

Alt 2 is preferable since it will have smaller variance, and less bias. Ok

b) Alt 1: Total sample size = $50 \cdot 9 = 450$ Ok

Alt 2: Total sample size = $60 \cdot 3 = 180$ Ok

(6)

(8) a) ... the problem ... is the 580 ... this ... steps. OK, accept / fix 2

One ... could ... hot deck imputation, we use the ... for ... small ... in this case we could use box 4

... method is to ... data without this lost value, and ... the ... set be 1 now OK 2

c) Using hot deck imputation $\hat{t}_y = \frac{N}{n}\sum_{i \in S} y_i = \frac{580}{12} \cdot 125 \approx 6041,66$

$\hat{V}(\hat{t}_y) \quad N^2\left(1-\frac{n}{N}\right)\frac{s_y^2}{n} = 580^2\left(1-\frac{12}{580}\right)\frac{32,8106}{12} \approx 900766,505$ SE 949,084

$s_y^2 \quad \frac{1}{11}\sum(y_i - \bar{y}_s)^2 = \frac{1}{11}\left((1438+225) - \frac{175^2}{12}\right) = 32,8106$ $6041,66 \pm 1,96 \cdot 949,084$

t-distribution b/c/ because small sample size. I don't expect an exact value, but some note

The estimate of the total is 6041,66, 95% CI: [4181,455; 7901,865] ok

Using 11 remaining obs: $\hat{t}_y = \frac{580}{11} \cdot 110 \approx 5800$ SE 007,006

$\hat{V}(\hat{t}_y) = 580^2\left(1-\frac{11}{580}\right)\frac{33,8}{11} = 1014061,4546$

$s_y^2 = \frac{1}{10}\left(1438 - \frac{12100}{11}\right) = 33,8$ $5800 \pm 1,96 \cdot 1007,006$

The estimate of the total is 5800, 95% CI: [3826,268; 7773,73] OK

The hot-deck imputation will give a negative bias on the variance if the true value was further from the mean than the imputed.

When disregarding the missing value, if got missing completely at random, the variance will remain biased. If the count was lost because it was very high for example the variance will have a negative bias 11

SU, DEPARTMENT OF STATISTICS

Room: BR Anonymous code: 0012-CDS Sheet number: 03

(9) $n = 31 \quad N = 2967 \quad t_0 = 1062$

a The ratio is calculated by: $\hat{B} = \frac{\hat{t}_y}{\hat{t}_x} = \frac{26,5}{10,4} = 2,548 \ m^3/m$ 04 1

b $\hat{t}_{HT} = \sum_{i \in S} \frac{y_i}{\pi_i} = \frac{26,5}{(31/2967)} = 2536,306$ 04

$\left(1 - \frac{31}{2967}\right) \approx 0,99$

$\hat{V}(\hat{t}_{HT}) = 2967^2 \cdot 0,99 \cdot \frac{0,22}{31} = 61848,799$

SE: 248,694 02 2

c $\hat{t}_{rat} = t_x \frac{\hat{t}_y}{\hat{t}_x} = 1062 \frac{26,5}{10,4} = 2706,058$ 04

$\hat{V}(\hat{t}_{rat}) = 2967^2 \cdot 0,99 \cdot \frac{s_e^2}{n}$

$= 2967^2 \cdot 0,99 \cdot \frac{0,559427}{31} = 157229,447$

SE: 396,520

155

$s_e^2 = s_y^2 + \hat{B}^2 s_x^2 - 2\hat{B} s_{xy}$

$s_{xy} = R \cdot s_x \cdot s_y = 0,94 \cdot 0,09 \cdot 0,47 = 0,035344$

$s_e^2 = 0,22 + 2,548^2 \cdot 0,08 - 2 \cdot 2,548 \cdot 0,035344$

$= 0,559427$ 0,006 1

0.08

d $\hat{t}_{reg} = N\left(\bar{y}_s + \hat{B}_1(\bar{x}_U - \bar{x}_s)\right) = 2967\left(\frac{26,5}{31} + 5,66\left(\frac{1062}{2967} - \frac{16,4}{31}\right)\right) = 2913,372$ 04

$\hat{V}(\hat{t}_{reg}) = N^2\left(1 - \frac{n}{N}\right)\frac{s_y^2}{n}(1 - R^2) = 2967^2 \cdot 0,99 \cdot \frac{0,22}{31} \cdot \left(1 - 0,94^2\right) = 7197,2$ 3

SE: 84,948 04

e The data clearly follows the fitted line, why the regression estimator has the lowest variance. It is also centered around the mean, explaining second place. HT. The ratio estimator is worst since it assumes that the data linearly approaches 0 on both variables, which is not the case. 1 8