Stockholms universitet

OBS! Läs noga igenom anvisningarna i tentamen, t.ex. hur du ska skriva svaren. Det är ditt ansvar som student att följa de anvisningar som ges.

NOTE! Read the examination instructions carefully, e.g. how to write the answers. It is your responsibility as a student to follow the given instructions.

Skriv din anonymiseringskod och dagens datum på allt material du lämnar in.
(*Enter your anonymization code and todays date on all submitted materials*)

Anonymiseringskod (*Anonymization code*)	3	1	1	-	0	0	0	6	-	D	A	S

Datum (*Date* YYYY-MM-DD)	2021-10-29	Plats nr. *(Seat No.)*	32

Kurs/Kurskod (*Course/Course code*)	ST306G
Kursmoment (*Course component*)	Survey sampling

Fylls i av tentamensvärd (*To be filled in by invigilator*)

Direkt i skrivning: (kryss)		Svarsblankett: (kryss)		Lösa svarsblad: (antal)	6

Lämnat in blankt: (kryss)		Dator: (kryss)	

Inlämningstid: 18:00 Signatur tentamensvärd: RR

Fylls i av lärare/examinator (*To be filled in by teacher/examinator*)

Betyg:	A	Poäng:	48

Signatur rättande lärare/examinator:

Regler i skrivsalen

- Följ tentamensvärds anvisningar.
- Väskor och ytterkläder ska placeras på anvisad plats.
- Placera ID-handling väl synlig på bordet framför dig.
- Ingen student får lämna skrivsalen under de första 30 minuterna.
- Endast en student i taget får besöka toaletten. Vid toalettbesök skriv ditt namn och klockslag på avsedd lista. Efter toalettbesöket ska du åter ange klockslag på listan.
- Elektronisk utrustning som mobiltelefon eller Smartwatch ska vara avstängd och placerad på anvisad plats.
- Under tentamen gäller tystnad – det är förbjudet att prata, eller på annat sätt kommunicera, med andra studenter under pågående tentamen.
- Innan tentamenshandlingarna lämnas in; skriv sidnummer, anonymiseringskod och datum på alla inlämnade papper.

Om något är oklart – fråga gärna tentamensvärden. Lycka till!

Rules in the examination hall

- Follow the invigilator's instructions.
- Bags and outerwear must be placed at the designated place.
- Place your ID document clearly visible on the table in front of you.
- No student may leave the examination hall for the first 30 minutes.
- Only one student at a time may visit the toilet. Before visiting the toilet, write your name and time on the intended list. After the toilet visit, enter the time on the list again.
- Electronic equipment such as a mobile phone or Smartwatch must be switched off and placed at the designated place.
- During the exam, silence applies – you are not allowed to talk, or otherwise communicate, with other students during the exam.
- Before submitting the examination documents; remember to write the page number, anonymization code, and date on all papers.

Please do not hesitate to ask the invigilator if anything is unclear. Good luck!

Tentamensservice
tentamensservice@su.se | www.su.se/tentamensservice

Stockholms universitet

Stockholms universitet

Datum: (*Date YYYY-MM-DD*)	2021-10-29	Kurs/Kurskod: (*Course/Course code*)	ST306G	Sidnr.: (*Page no.*)	1
Anonymiseringskod (*Anonymization code*)	311-0006-DAS				

Uppg.nr.: (*Task no.*) 1-5

Lärarens kommentar: (*Teacher's note*)

1. A. R

2. ~~b.~~ d

3. C. R

4 E. R

5. C. R

Poäng: (*Points*) 8

Tentamensservice/SU/typ_a

Uppg.nr.:
(*Task no.*)

Lärarens kommentar:
(*Teacher's note*)

Poäng:
(*Points*)

a./ Because yield is a variable of farms, like income is for persons, and we want to estimate total yield in country X of runner beans and other crops, our target population is all active farms in country X. This assuming only farms grow crops. good point. If one were to be philosophical one could argue that the target population is the runner beans that have been grown in country X and that their weight is the variable of interest. But I think the first definition is much more practical and realistic. 2

b./ Primary sampling units are the farms in the frame for method 2. Secondary sampling units are the plots. 1

c./ for method 1, the observation units are the farms that were sampled from the frame.

for method 2, the observation units are the plots of land sampled at the farms. Here the weight of the crops would be a variable of the plot. 2

Poäng: (Points)

Uppg.nr.:
(*Task no.*)

Lärarens kommentar:
(*Teacher's note*)

Poäng:
(*Points*)

d./ The first method would probably be subject to non-response, with farms with low yield not bothering to answer the questionnaire because they got worse things to deal with. This could introduce non-response bias. But answers could be pretty precise as farms must keep track of their yields at a very exact scale.

Yes true. But the survey request may have come before they knew

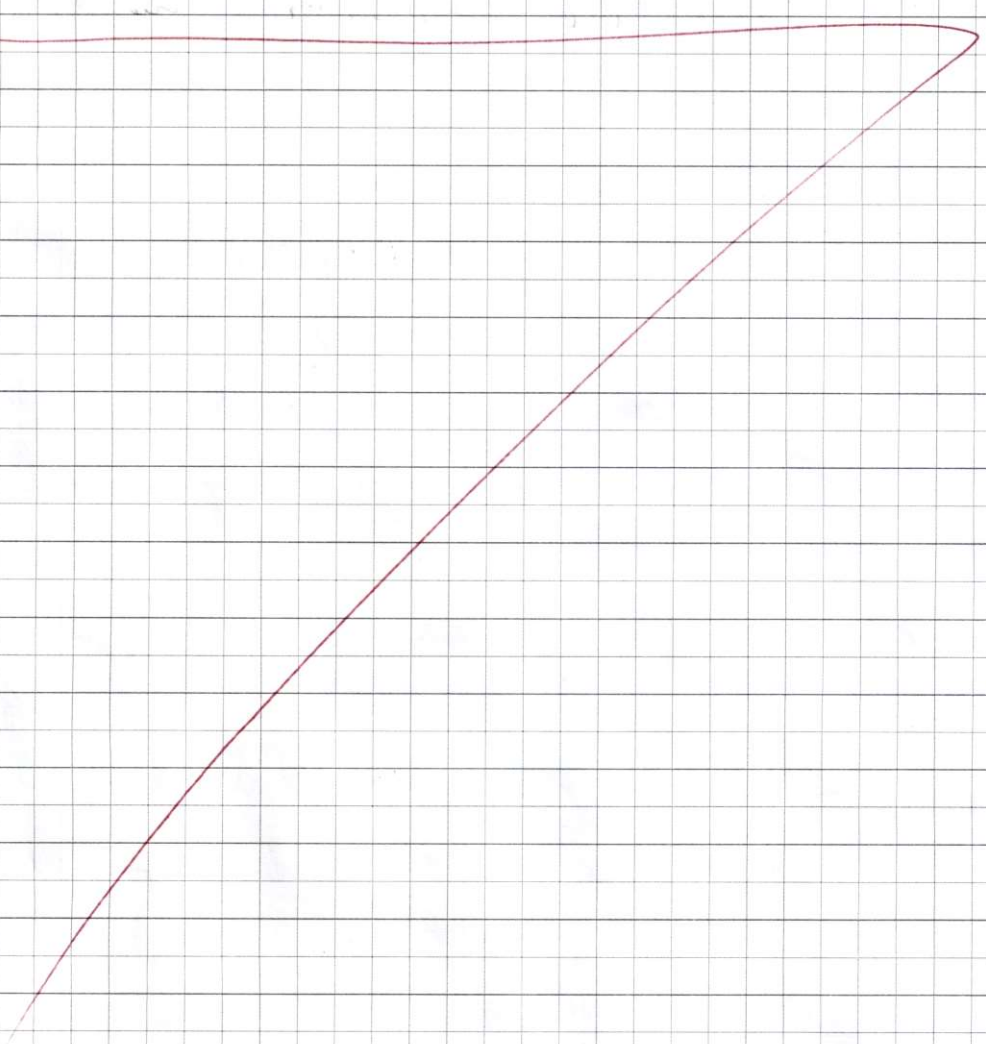
The second method could be more imprecise as only 2 plots are sampled at each farm, making the estimate potentially unstable. Also the persons sent may not be as experienced as the farmers themselves at measuring yield, which may cause measurement bias.

The second method could be seen as an 2-stage cluster sample with only two SSU sampled, which is not great! I would probably choose method 1 and trust the farmers precision and write that they gain no subsidies by underreporting yield in the survey so as to minimize lying. I would adjust for non-response best I can after. And design the survey with non-response in mind. Turn!

5

e./ I would use weighting class estimator and use the indicator variable as the class determiner. I think this would be preferable to just using the regular HT estimator as we are bound to have some non-response. I would assume MAR and if it would turn out to be so the weighting class estimator would probably be an improvment over HT-estimator.

2



Uppg.nr.: (*Task no.*) 6.

Lärarens kommentar: (*Teacher's note*)

Poäng: (*Points*) 12

Uppg.nr.: (Task no.) 7.

Lärarens kommentar: (Teacher's note)

7. A stratified simple random sample. R

Each section is a stratum and two SRS are drawn independently. Moreover it has a disproportional allocation. ↑

b. $y = \begin{cases} 1 & \text{if woman} \\ 0 & \text{otherwise} \end{cases}$ $\bar{y} = \hat{p}$ $N = 2500$

survey association →

$\hat{t}_{survey} = \sum y_i \cdot \frac{N_{survey}}{n_{survey}} \rightarrow 80 \cdot \frac{2000}{200} = \underline{800}$

$\hat{p}_{survey} = \frac{\sum y_i}{n_{survey}} \rightarrow \frac{80}{200} = \underline{0,4}$ R 1

the cramer society →

$\hat{t}_{cramer} = \sum y_i \frac{N_{cramer}}{n_{cramer}} \rightarrow 40 \cdot \frac{500}{200} = \underline{100}$

$\hat{p}_{cramer} = \frac{\sum y_i}{n_{cramer}} \rightarrow \frac{40}{200} = \underline{0,2}$ R 1

The whole statistical society →

$\hat{t}_{str} = \sum \hat{t} \rightarrow 800 + 100 = \underline{900}$

$\hat{p}_{str} = \frac{\hat{t}_{str}}{N} \rightarrow \frac{900}{2500} = \underline{0,36}$ R 2

$\hat{V}(\hat{p}_{str}) = \frac{1}{N^2} \sum^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}$, $s_h^2 \approx \hat{p}_h(1-\hat{p}_h)$

correct (

$\hat{V}(\hat{p}_{str}) = \left(2000^2 \left(1 - \frac{200}{2000}\right) \frac{0,4(1-0,4)}{200}\right) + \left(500^2 \left(1 - \frac{200}{500}\right) \frac{0,2(1-0,2)}{200}\right) \Big/ 2500^2 = \underline{0,0007104}$

correct same mistake?

TURN!

Poäng: (Points)

Uppg.nr.: (*Task no.*) 7.

Lärarens kommentar: (*Teacher's note*)

c./ yes it is fair to say that survey association has a greater proportion of women than the Cramer society.

If we look at the estimates we see great difference → $\hat{p}_{survey} = 0,4$

$\hat{p}_{Cramer} = 0,2$

lets And $\hat{SE}(\hat{p}_A - \hat{p}_B)$

$$\sqrt{\left(1-\frac{200}{2000}\right)\frac{0,4(1-0,4)}{200} + \left(1-\frac{200}{500}\right)\frac{0,2(1-0,2)}{200}} = 0,0394 \text{ R}$$ 2

$2 \cdot 0,0394 = 0,07899$.

The estimated difference's CI does not contain 0 → strong indication of our conclusion! Good

d./ Domain estimation →

$\bar{y}_d = \frac{\bar{u}_s}{\bar{x}_s} = \frac{\bar{t}_u}{\bar{t}_x}$ where $x =$ 1 if new, 0 otherwise; $u = y$ if new, 0 otherwise

mean new members → $\frac{28000 + 35000 + 36000}{3} = \frac{33'000}{3} = \bar{y}_d$ R

variance → $\hat{V}(\bar{y}_d) = \left(1-\frac{n}{N}\right)\frac{s_{y_d}^2}{n_d}$ $\quad s_{y_d}^2 = \frac{20'000}{2} = 10'000$

$\hat{V}(\bar{y}_d) = \left(1-\frac{200}{2000}\right)\frac{10'000}{3} = \frac{3000}{3}$ R

↑ because the survey association

Poäng: (*Points*) 12

$y = 1$ Interested, 0 not interested $\quad \bar{y} = p$, $p = 0,3$

Margin of error = 0,02

a./ Sample size for 95 % confidence interval and $e = 0,02$. Ignoring FPC →

$n = \frac{1,96^2 \cdot S_y^2}{e^2}$, $S_y^2 \approx p(1-p) = 0,3$

$n = \frac{1,96^2 \cdot 0,3(1-0,3)}{0,02^2}$, $n = 2016,84 \rightarrow n = 2017$ R 3

b./ Max $S_y^2 \rightarrow$ max $p(1-p)$, where $p \in (0,1) \rightarrow$

$p = 0,5 \rightarrow S_y^2 = 0,5(1-0,5) = 0,25$

Safe sample size →

$n = \frac{1,96^2 \cdot 0,25}{0,02^2}$, $n = 2401$ R 1

c./ Assumptions: MCAR → The response rate and response propensity R does not depend on y, x or sampling design. The response set is seen as a SRS subsample of the sample set. We can therefor ignore non-response. This assumtion is often not very realistic!

The sample mean is the estimated population mean → $\bar{y}_S = \hat{p} = 0,3$ = estimate for all customers

I also assume that respondents tell the truth! No non-sampling errors! measurement error! 3

Turn! →

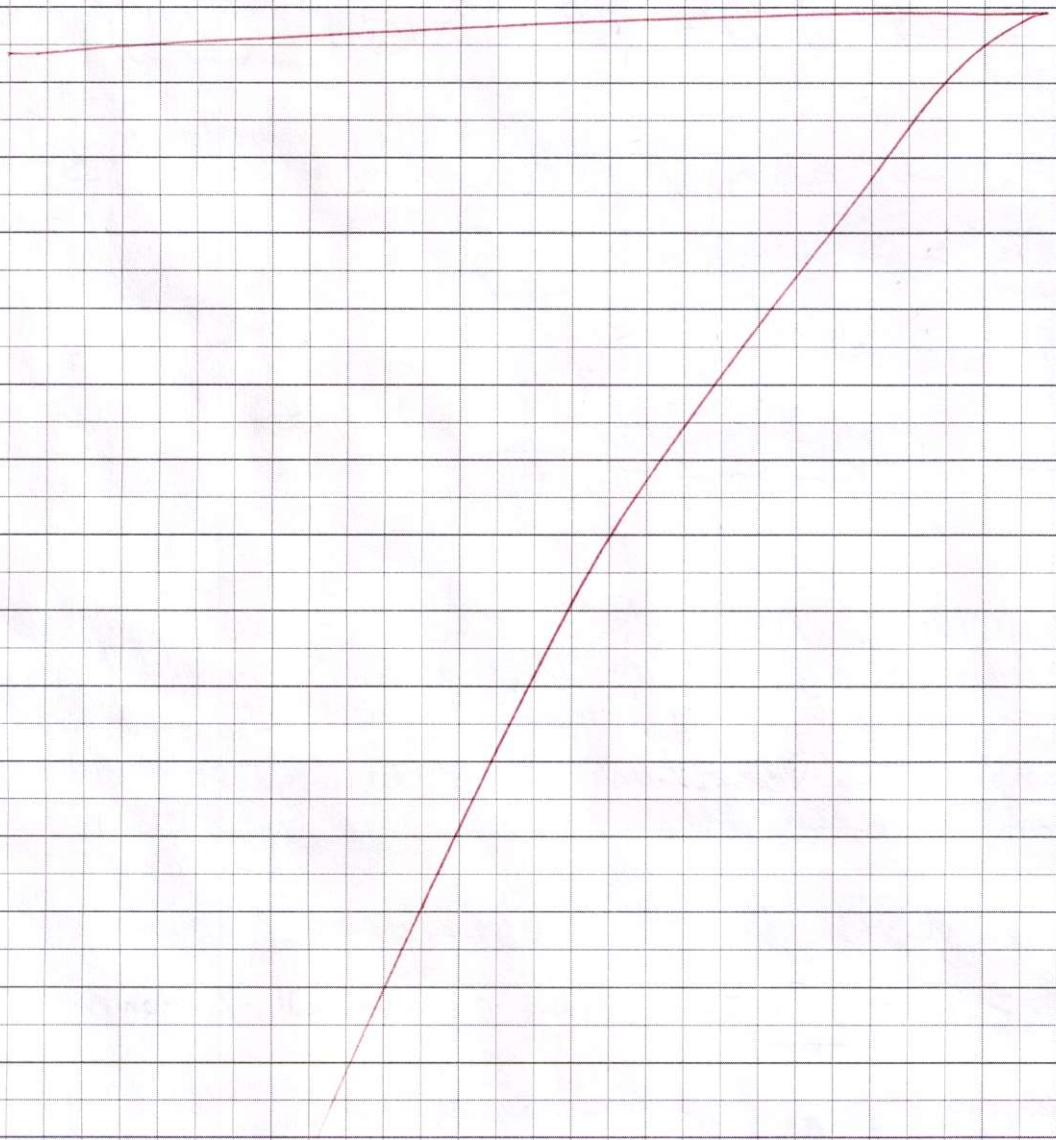
Poäng: (Points)

for 50% response

Uppg.nr.: (*Task no.*) 8.

Lärarens kommentar: (*Teacher's note*)

d) Possible reasons could be choice of mode, (email and telephone surveys can have lower response rate compared to in-person), poor reputation of agency (i.e Not a strong Athority) and bad presentation of the survey and Question (for example no personalization with the name of the sample unit in the email or greeting on the phone. Liking can play a part in response rate. 3



Poäng: (*Points*) 10

9./ First estimator → HT estimator (unbiased)

$\hat{t}_y = \sum y_i \frac{N}{n} \rightarrow \bar{y} \cdot N \quad N = 1000 \quad n = 10$

$\hat{t}_y = 5{,}38 \cdot \frac{1000}{10} = 538$ R

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

$\hat{V}(\hat{t}_y) = 1000^2 \left(1 - \frac{10}{1000}\right) \frac{0{,}08}{10} = 7920$ R

Second estimator → Regression estimator (approx. unbiased)

$\hat{t}_{reg} = \hat{t}_y + \hat{B}_1 (t_x - \hat{t}_x), \quad \hat{B}_1 = \frac{s_{yx}}{s_x^2} = \frac{\sum (y_k - \bar{y})(x_k - \bar{x})}{\sum (x_k - \bar{x})^2}$

$\hat{B}_1 = \frac{0{,}13}{0{,}54} = 0{,}2407$

$t_x = 447, \quad \hat{t}_x = 1{,}47 \cdot \frac{1000}{10} = 447$

$\hat{t}_{reg} = 538 + (0{,}2407)(447 - 447) = 538$ R

$\hat{V}(\hat{t}_{reg}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n} (1 - R^2), \quad R^2 = \left(\frac{s_{yx}}{s_y s_x}\right)^2$

$\hat{V}(\hat{t}_{reg}) = 1000^2 \left(1 - \frac{10}{1000}\right) \frac{0{,}08}{10} (1 - (0{,}216)^2) = 7550{,}4844$ i2

4

b/ looking at the graph, the y values from the new sample looks about the same, so the HT estimator's estimate would not change much. The x values are vastly different although we seem to have gotten a "bad sample" if considering the x values, compared to the first sample when the estimator was exactly right (for x). So the regression correction term would in this case be different (by quite a lot)! This would result in a much different estimate for $\bar{t}_y$ when using regression estimator. I would rank HT estimator better than regression estimator in this case.

Poäng: (Points)

Turn! →

Uppg.nr.: (*Task no.*) 9.

Lärarens kommentar: (*Teacher's note*)

b. cont.

I would rank HT better in this case as we now have two samples of y values giving similar HT estimates so we are getting good information that $\hat{t}_y$ is around 530 or so. To choose the regression estimator in the second case would be to let the "bad sample" of x have to much weight in my opinion! Not sure I would agree, but I accept your line of argument.

Poäng: (*Points*) 6