STOCKHOLMS UNIVERSITET
Statistiska institutionen
Michael Carlson (examinator)

HT2019

Tentamen i Undersökningsmetodik (4,5 hp)

Kurs: Regressionsanalys och undersökningsmetodik

2020-01-15

Skrivtid: kl. 16.00 - 21.00 (5 timmar)
Godkända hjälpmedel: Miniräknare utan lagrade formler och text
Vidhäftade hjälpmedel: Formelsamling och Statistiska tabeller (endast de tabeller som krävs)

- Tentamen består av 5 uppgifter, i förekommande fall uppdelade i deluppgifter. Maximalt antal poäng anges per deluppgift.
- Svar med fullständiga redovisningar ska lämnas.
  - Använd endast skrivpapper som tillhandahålls i skrivsalen.
  - För full poäng på en uppgift krävs tydliga, utförliga och väl motiverade lösningar.
  - Kontrollera alltid dina beräkningar och lösningar! Slarvfel kan också ge poängavdrag!
  - Använd minst fem värdesiffror i dina beräkningar (1,2345 och 1234,5 är exempel på tal med fem värdesiffror). I förekommande fall är det inte möjligt pga. avrundning i t.ex. SAS-utskrifter men utgå då ifrån det som är givet. Du kan dock avrunda ditt slutliga svar.
- Tentamen kan maximalt ge 100 poäng och för godkänt resultat krävs minst 50.
- Betygsgränser:

  A: 90 – 100 p
  B: 80 – 89 p
  C: 70 – 79 p
  D: 60 – 69 p
  E: 50 – 59 p
  Fx: 40 – 49 p
  F: 0 – 40 p

  OBS! Fx och F är underkända betyg som kräver omexamination. Studenter som får betyget Fx kan alltså inte komplettera för högre betyg.
- Lösningsförslag läggs ut på Athena kort efter tentamen.

LYCKA TILL!

Uppgift 1. (20p)

En utredning genomfördes i en liten ort av kommunkontoret för att skatta P = andelen hushåll med minst en hushållsmedlem över 65 år gammal. På orten fanns totalt 720 hushåll. Ett obundet slumpmässigt urval (utan återläggning) av storlek 60 drogs från populationen av hushåll. Av dessa fann man att 12 av de observerade hushållen hade minst en hushållsmedlem som var över 65 år gammal, i alla övriga hushåll var samtliga medlemmar under 65 år.

a) Skatta antalet hushåll som har minst en medlem över 65 år och beräkna ett 95 % konfidensintervall för skattningen. (10p)

Samma undersökning skulle senare genomföras i en grannkommun med totalt 600 hushåll. Här var man helt okunnig om hushållens sammansättning men misstänkte att den sanna andelen var större än i den första undersökningen.

b) Bestäm minsta stickprovsstorkleken för den nya undersökningen givet att man vill få ett resultat med minst samma precision som i första undersökningen. Använd formeln i formelsamlingen men tänk på att du måste ersätta σ^2 med något annat. Vad? TIPS: Jämför formlerna för $V(\bar{y})$ och $V(\hat{p})$. (10p)

Uppgift 2. (30p)

Ett företag studerar Y = frånvaron räknat i timmar bland de anställda på en monteringsfabrik under en given månad där orsakerna till frånvaron kunde relateras till enklare arbetsskador. De anställda kan grovt delas upp i tre grupper: montörer, tekniker/ingenjörer och administrativ personal. Eftersom skaderiskerna för dessa tre grupper ser olika ut beroende på olikheter i arbetsmiljö så bestämmer man sig för att dra ett stratifierat urval med OSU utan återläggning inom varje stratum. Man vet hur många anställda som finns inom respektive stratum och från en motsvarande undersökning från året innan får man ungefärliga siffor för de tre gruppernas varianser:

1. Montörer	2. Tekniker/ingenjörer	3. Administratörer
$\sigma_1^2 = 100$	$\sigma_2^2 = 64$	$\sigma_3^2 = 25$
$N_1 = 132$	$N_2 = 92$	$N_3 = 27$

a) Bestäm en lämplig stickprovsallokering för ett stickprov med totalt $n = 30$. (10p)

Man drar ett stratifierat urval och får resultaten som återges i tabellen på nästa sida. Notera att tabellen även återger stickprovsmedelvärdet och stickprovsvariansen ovägt, dvs. ej uppdelat på strata utan för hela stickprovet.

b) Skatta den genomsnittliga sjukfrånvaron per anställd och beräkna ett 95 % konfidensintervall utifrån den stratifierade urvalsdesignen. (10p)

c) Skatta den genomsnittliga sjukfrånvaron per anställd och beräkna ett 95 % konfidensintervall som om det vore ett enkelt OSU. (7p)

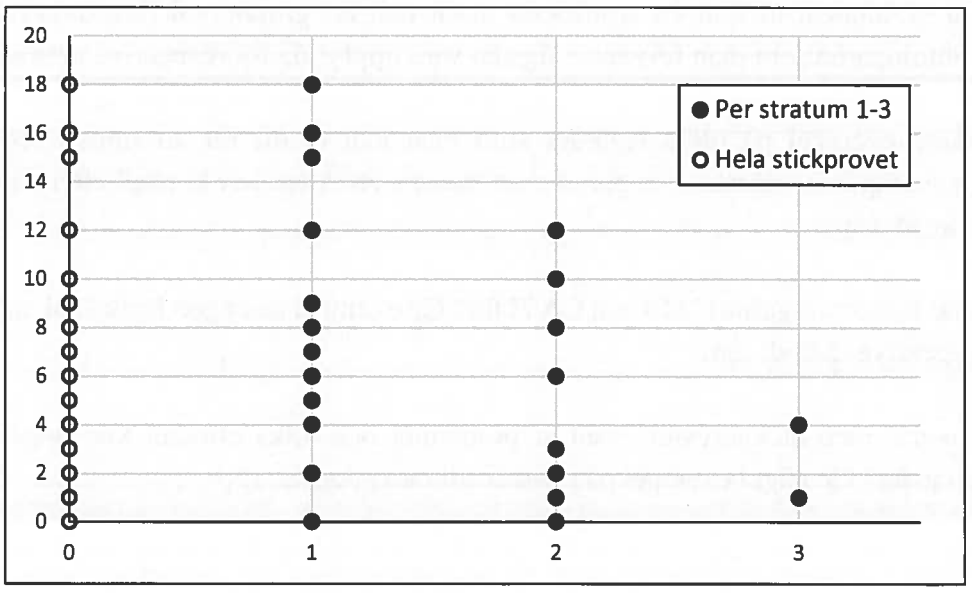
(forts. nästa sida)

(forts. Uppgift 2)

d) Kan du ge en enkel förklaring för varför skillnaden mellan felmarginalerna inte är så stor i detta fall? TIPS: Studera varianserna samt diagrammet nedan. Se även fråga 4a nedan. (3p)

Stratum 1		Stratum 2		Stratum 3	
0	0	0	0	1	4
0	0	1	2		
2	4	3	6		
4	5	8	8		
6	7	10	12		
8	8				
9	12				
15	16				
16	18				
$N_1 = 132$		$N_2 = 92$		$N_3 = 27$	
$n_1 = 18$		$n_2 = 10$		$n_3 = 2$	
$\bar{y}_1 = 7.2222$		$\bar{y}_2 = 5$		$\bar{y}_3 = 2.5$	
$s_1^2 = 36.536$		$s_2^2 = 19.111$		$s_3^2 = 4.5$	
Hela stickprovet totalt (ovägt):					
$N = 251$	$n = 30$	$\bar{y} = 6.1667$		$s^2 = 29.5920$	

Fördelningen för y för stickprovet:



Uppgift 3. (20p)

Efter en stor marknadsföringskampanj har försäljningsdata för en viss produkt samlats in under en månad för $n = 13$ försäljningsställen ur en population av $N = 104$. Urvalet av försäljningsställen gjordes med ett enkelt OSU utan återläggning. Förra årets försäljning för samma månad finns tillgängliga för samtliga försäljningsställen.

Låt y beteckna årets månadsförsäljning (efter kampanjen) och låt x beteckna förra årets månadsförsäljning (före kampanjen), båda i tkr. Förra årets totala försäljning för samtliga försäljningsställen var $\tau_x = 1144$ tkr. Följande summor finns nu tillgängliga:

$$\sum_{k\in S} x_k = 130 \qquad \sum_{k\in S} y_k = 182$$

$$\sum_{k\in S} x_k^2 = 1444 \qquad \sum_{k\in S} y_k^2 = 2865 \qquad \sum_{k\in S} x_k y_k = 2020$$

a) Skatta τ_y = den totala månadsförsäljningen under detta år genom att använda en kvotestimator med förra årets månadsförsäljning som hjälpvariabel. Beräkna sedan standardfelet för $\hat{\tau}_{\text{kvot}}$. (10p)

b) Jämför skattningen i a) med den vanliga HT-skattningen under OSU. Beräkna punktskattningen $N\bar{y}$ och motsvarande standardfel och kommentera kortfattat. Lönade det sig att använda kvotskattningen eller inte? (10p)

Uppgift 4. (20p)

För var och en av följande deluppgifter ska du svara kortfattat. Hela uppgiften bör kunna redovisas på maximalt ca två A4-sidor. Du får gärna komplettera med bilder och skisser.

a) Vad är skillnaden mellan ett stratifierat urval och ett gruppurval (klusterurval)? Vilka är förutsättningarna som man förväntar sig ska vara uppfyllda för respektive urvalsdesign? (5p)

b) Ge några exempel på olika åtgärder som man kan ta till för att minska bortfallet i en undersökning och vad man kan göra för att minska effekterna av bortfall efter att insamlingen är avslutad. (5p)

c) Vad står förkortningarna CAPI och CATI för? Ge exempel på någon fördel och någon nackdel för respektive metod. (5p)

d) Vad menas med täckningsfel? Vad är problemet och vilka effekter kan uppstå om det är täckningsfel? Ge några exempel på orsaker till täckningsfel. (5p)

Uppgift 5. (10p)

Man har en population bestående av $N = 5$ element med följande värden på en variabel Y:

$$U = \{2, 5, 6, 11, 16\}$$

a) Lista alla möjliga stickprov med storlek $n = 2$ som kan dras från U med OSU utan återläggning. Beräkna sedan stickprovsmedelvärdet $\bar{y}$ för vart och ett av dessa stickprov. KOM IHÅG: Det finns $\binom{N}{n}$ olika stickprov så du ska få lika många stickprovsmedelvärden. (3p)

b) Beräkna μ för population och sedan $\bar{\bar{y}}$ = medelvärdet av medelvärdena som du fick i a). Kan du för detta fall bekräfta att $\bar{y}$ är en väntevärdesriktig skattning för μ? (3p)

c) Beräkna den teoretiska variansen $V(\bar{y})$ för stickprovsmedelvärdet och sedan $\sigma_{\bar{y}}^2$ = variansen för stickprovsmedelvärdena i a). Ska dessa två varianser vara lika stora? (4p)

Formel- och tabellsamling

DESKRIPTIV STATISTIK

Notation: U = populationen
S = stickprov (stort S); $\subseteq U$

Medelvärde:	$\mu = \frac{1}{N}\sum_{k\in U} y_k$	Varians:	$\sigma^2 = \frac{\sum_{k\in U}(y_k-\mu_y)^2}{N} = \frac{\sum_{k\in U} y_k^2 - N\mu_y^2}{N}$
	$\bar{y} = \frac{1}{n}\sum_{k\in S} y_k$		$s^2 = \frac{\sum_{k\in S}(y_k-\bar{y})^2}{n-1} = \frac{\sum_{k\in S} y_k^2 - n\bar{y}^2}{n-1}$
Andel:	$P = \frac{1}{N}\sum_{k\in U} y_k$		$\sigma^2 = P(1-P)$
($y_k = 0$ eller 1)	$\hat{p} = \frac{1}{n}\sum_{k\in S} y_k$		$s^2 = \frac{n}{n-1}\hat{p}(1-\hat{p})$
Kovarians:	$\sigma_{xy} = Cov(x,y) = \frac{\sum_{k\in U}(x_k-\mu_x)(y_k-\mu_y)}{n-1} = \frac{\sum_{k\in U} x_i y_i - n\bar{x}\bar{y}}{n-1}$		
	$s_{xy} = Cov(x,y) = \frac{\sum_{k\in U}(x_k-\bar{x})(y_k-\bar{y})}{n-1} = \frac{\sum_{k\in U} x_k y_k - n\bar{x}\bar{y}}{n-1}$		
Korrelation:	$r_{xy} = Corr(x,y) = \frac{s_{xy}}{s_x \cdot s_y} = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}}$		

Beräkningsformler för VARIANSER och REGRESSIONSKOEFFICIENT

$$s^2 = \frac{n\sum y_k^2 - (\sum y_k)^2}{n(n-1)} = \frac{\sum y_k^2 - \frac{(\sum y_k)^2}{n}}{n-1} = \frac{\sum y_k^2 - n\bar{y}^2}{n-1} = \frac{\sum(y_k-\bar{y})^2}{n-1}$$

$$b = \frac{n\sum x_k y_k - (\sum x_k)(\sum y_k)}{n\sum x_k^2 - (\sum x_k)^2} = \frac{\sum x_k y_k - \frac{(\sum x_k)(\sum y_k)}{n}}{\sum x_k^2 - \frac{(\sum x_k)^2}{n}} = \frac{\sum x_k y_k - n\bar{x}\bar{y}}{\sum x_k^2 - n\bar{x}^2}$$

$$= \frac{\sum(x_k-\bar{x})(y_k-\bar{y})}{\sum(x_k-\bar{x})^2} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})/(n-1)}{\sum(x_i-\bar{x})^2/(n-1)}$$

$$= \frac{s_{xy}}{s_x^2} = \frac{s_{xy}}{s_x^2}\cdot\frac{s_x s_y}{s_x s_y} = \frac{s_{xy}}{s_x s_y}\cdot\frac{s_y}{s_x} = r_{xy}\cdot\frac{s_y}{s_x}$$

OBS! Notationen har förenklats ovan, summationsindex är alltid k: ex. $\sum y_k = \sum_{k\in S} y_k$

OBUNDET SLUMPMÄSSIGT URVAL u.å.

Parameter:	Estimator:	Varians $V(\cdot)$:	Variansskattning $\hat{V}(\cdot)$:
μ	$\bar{y} = \frac{1}{n}\sum_{k\in S} y_k$	$V(\bar{y}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$	$\hat{V}(\bar{y}) = \left(1-\frac{n}{N}\right)\frac{s^2}{n}$
τ	$\hat{t} = N\bar{y}$	$V(\hat{t}) = N^2 V(\bar{y})$	$\hat{V}(\hat{t}) = N^2 \cdot \hat{V}(\bar{y})$
P	$\hat{p} = \frac{1}{n}\sum_{k\in S} y_k$	$V(\hat{p}) = \left(\frac{N-n}{N-1}\right)\frac{P(1-P)}{n}$	$\hat{V}(\hat{p}) = \left(1-\frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}$
A	$\hat{A} = N\hat{p}$	$V(\hat{A}) = N^2 V(\hat{p})$	$\hat{V}(\hat{A}) = N^2 \cdot \hat{V}(\hat{p})$

Stickprovsstorlek: $$n \geq \frac{N\sigma^2}{D^2(N-1)+\sigma^2}$$

STRATIFIERAT URVAL u.å.

Notation: L = antal strata

N_k = populationsstorleken för stratum $k = 1, \dots, L$

n_k = stickprovets storlek i stratum $k = 1, \dots, L$

$W_k = N_k/N$

$\bar{y}_k$ = stickprovsmedelvärde i stratum $k = 1, \dots, L$

s_k^2 = stickprovsvarians i stratum $k = 1, \dots, L$

Parameter	Estimator	Varians $V(\cdot)$	Variansskattning $\hat{V}(\cdot)$
μ	$\bar{y}_{\text{str}} = \sum_{k=1}^{L} W_k \bar{y}_k$	$\sum_{k=1}^{L} W_k^2 \left(\frac{N_k-n_k}{N_k-1}\right)\frac{\sigma_k^2}{n_k}$	$\sum_{k=1}^{L} W_k^2 \left(1-\frac{n_k}{N_k}\right)\frac{s_k^2}{n_k}$
τ	$\hat{t}_{\text{str}} = N\bar{y}_{\text{str}}$	$\sum_{k=1}^{L} N_k^2 \left(\frac{N_k-n_k}{N_k-1}\right)\frac{\sigma_k^2}{n_k}$	$\sum_{k=1}^{L} N_k^2 \left(1-\frac{n_k}{N_k}\right)\frac{s_k^2}{n_k}$
P	$\hat{p}_{\text{str}} = \sum_{k=1}^{L} W_k \hat{p}_k$	$\sum_{k=1}^{L} W_k^2 \left(\frac{N_k-n_k}{N_k-1}\right)\frac{P_k(1-P_k)}{n_k}$	$\sum_{k=1}^{L} W_k^2 \left(1-\frac{n_k}{N_k}\right)\frac{\hat{p}_k(1-\hat{p}_k)}{n_k-1}$
A	$\hat{A}_{\text{str}} = N\hat{p}_{\text{str}}$	$\sum_{k=1}^{L} N_k^2 \left(\frac{N_k-n_k}{N_k-1}\right)\frac{P_k(1-P_k)}{n_k}$	$\sum_{k=1}^{L} N_k^2 \left(1-\frac{n_k}{N_k}\right)\frac{\hat{p}_k(1-\hat{p}_k)}{n_k-1}$

Optimal allokering: $$n_k = n \cdot \frac{N_k\sigma_k}{\sum_{j=1}^{L} N_j\sigma_j}$$

KLUSTERURVAL – OSU u.å.

Notation:

U = population av kluster
S = stickprov av kluster
N = antal kluster totalt
n = antal kluster i stickprovet

M = totalt antal element
m_i = antal element i kluster nr $i = 1,2 \ldots, N$
$\bar{m}$ = stickprovsmedelvärde av klusterstorlekarna m_i
$s^2_{m_i}$ = stickprovsvariansen av klusterstorlekarna m_i

$\tau = \sum_{k \in U} y_k$ = totalvärdet för y i hela populationen
$\mu = \tau / M$ = populationsmedelvärde av y
$\tau_i = \sum_{k \in C_i} y_k$ = totalvärdet för kluster nr $i = 1,2 \ldots, N$
$\bar{\tau}$ = stickprovsmedelvärde av totalvärdena τ_i
$s^2_{\tau_i}$ = stickprovsvariansen av totalvärdena τ_i

$A = \sum_{k \in U} y_k$ = antalet ettor i hela populationen; ($y_k = 0$ eller 1)
$P = A/M$ = andelen ettor i hela populationen; ($y_k = 0$ eller 1)

Parameter	Estimator	Variansskattning
M	$\hat{M}_{\text{vvr}} = N \cdot \bar{m}$	$\hat{V}(\hat{M}_{\text{vvr}}) = N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{s^2_{m_i}}{n}$
μ	$\bar{y}_{\text{vvr}} = \frac{\hat{t}_{\text{vvr}}}{M} = \frac{N\bar{\tau}}{M}$	$\hat{V}(\bar{y}_{\text{vvr}}) = \frac{N^2}{M^2} \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{s^2_{\tau_i}}{n}$
	$\bar{y}_{\text{kvot}} = \frac{\hat{t}_{\text{vvr}}}{\hat{M}} = \frac{\sum_{i \in S} \tau_i}{\sum_{i \in S} m_i}$	$\hat{V}(\bar{y}_{\text{kvot}}) = \left(\frac{1}{\bar{m}}\right)^2 \left(1 - \frac{n}{N}\right) \frac{\sum_{i \in S} (\tau_i - \bar{y}_{kvot} m_i)^2}{n(n-1)}$
τ	$\hat{t}_{\text{vvr}} = N\bar{\tau}$	$\hat{V}(\hat{t}_{\text{vvr}}) = N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{s^2_{\tau_i}}{n}$
	$\hat{t}_{\text{kvot}} = M\bar{y}_{\text{kvot}} = \frac{M}{\hat{M}} \hat{t}_{vvr}$	$\hat{V}(\hat{t}_{\text{kvot}}) = \left(\frac{M}{\bar{m}}\right)^2 \left(1 - \frac{n}{N}\right) \frac{\sum_{i \in S} (\tau_i - \bar{y}_{kvot} m_i)^2}{n(n-1)}$
P	*formler utgår*	
A	*formler utgår*	

SKATTNINGSMETODER

Notation: τ_y = totalvärdet för variabeln y för hela populationen
$\hat{\tau}_y$ = skattningen av τ_y under OSU
μ_y = populationsmedelvärdet av för variabeln y

Kvotskattning under OSU u.å.:

Parameter	Punkt- resp. variansskattning
τ_y	$\hat{\tau}_{\text{kvot}} = \hat{R} \cdot \tau_x = \frac{\sum_{k\in S} y_k}{\sum_{k\in S} x_k} \cdot \tau_x = \frac{\tau_x}{\hat{\tau}_x} \cdot \hat{\tau}_y$ där $\hat{R} = \frac{\sum_{k\in S} y_k}{\sum_{k\in S} x_k} = \frac{\hat{\tau}_y}{\hat{\tau}_x}$ $\hat{V}(\hat{\tau}_{\text{kvot}}) = N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{1}{n} \cdot \left(\frac{\sum_{k\in S}(y_k - \hat{R}x_k)^2}{n-1}\right)$ där $\sum_{k\in S}(y_k - \hat{R}x_k)^2 = \sum_{k\in S} y_k^2 - 2\hat{R}\sum_{k\in S} x_k y_k + \hat{R}^2 \sum_{k\in S} x_k^2$
μ_y	$\hat{\mu}_{\text{kvot}} = \hat{R} \cdot \mu_x = \frac{\sum_{k\in S} y_k}{\sum_{k\in S} x_k} \cdot \mu_x = \frac{\mu_x}{\bar{x}} \cdot \bar{y}$ $\hat{V}(\hat{\mu}_{\text{kvot}}) = \left(1 - \frac{n}{N}\right) \cdot \frac{1}{n} \cdot \left(\frac{\sum_{k\in S}(y_k - \hat{R}x_k)^2}{n-1}\right)$

Regressionsskattning under OSU u.å.:

Parameter	Punkt- och variansskattning
μ_y	$\hat{\mu}_{\text{reg}} = \bar{y} + b(\mu_x - \bar{x})$ där $b = \frac{\sum_{k\in S}(y_k - \bar{y})(x_k - \bar{x})}{\sum_{k\in S}(x_k - \bar{x})^2}$ $\hat{V}(\hat{\mu}_{\text{reg}}) = \left(1 - \frac{n}{N}\right) \cdot \frac{1}{n} \cdot \left(\frac{\sum_{k\in S}(y_k - \bar{y})^2 - b^2 \sum_{k\in S}(x_k - \bar{x})^2}{n-2}\right)$ där $\sum_{k\in S}(y_k - \bar{y})^2 = \sum_{k\in S} y_k^2 - n\bar{y}^2$
τ_y	$\hat{\tau}_{\text{reg}} = N \cdot \hat{\mu}_{\text{reg}}$ $\hat{V}(\hat{\tau}_{\text{reg}}) = N^2 \cdot \hat{V}(\hat{\mu}_{\text{reg}})$

Poststratifiering under OSU u.å.:

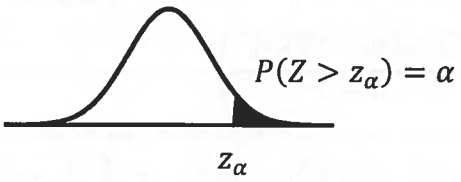
Parametrar och estimatorer - se under **Stratifierat urval** ovan
OBS! Populationsvikterna W_k måste vara kända.
Variansskattning – *formler utgår*

Från tabellsamlingen

TABELL 2. Normalfördelningens kvantiler, standardiserad

$Z \in N(0, 1)$. Vilket värde har z_α om $P(Z > z_\alpha) = \alpha$ där α är en given sannolikhet.

Utnyttja även $\Phi(-z) = 1 - \Phi(z)$ för $P(Z \leq -z_\alpha)$.



α	z_α
0,25	0,6745
0,10	1,2816
0,05	1,6449
0,025	1,9600
0,010	2,3263
0,005	2,5758
0,0025	2,8070
0,0010	3,0902
0,0005	3,2905
0,00025	3,4808
0,00010	3,7190
0,00005	3,8906
0,000025	4,0556
0,000010	4,2649
0,000005	4,4172

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 15/1-2020

Sal: Värtasalen

Tenta: Undersökningsmetodik

Kurs: Regressionsanalys och undersökningsmetodik

ANONYMKOD: 0068-ARH

☒ Jag godkänner att min tenta får läggas ut anonymt på hemsidan som studentsvar.

OBS! SKRIV ÄVEN PÅ BAKSIDAN AV SKRIVBLADEN

Markera besvarade uppgifter med kryss

1	2	3	4	5	6	7	8	9	Antal inl. blad
X	X	X	X	X					5
Lär.ant. 20	30	18	20	8					

RR

POÄNG	BETYG	Lärarens sign.
96	A	RR

SU, STATISTIK

Skrivsal: VÄ Anonymkod: 0068-ARH Blad nr: 1

1.

a) $N = 720 \quad n = 60 \quad$ OSU u. å.

$\hat{p} = \frac{12}{60} = 0,2$

$\hat{A} = N \cdot \hat{p} \qquad \hat{A} = 720 \cdot 0,2 = 144$ R

$\hat{V}(\hat{A}) = N^2 \cdot \hat{V}(\hat{p})$

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

$\hat{V}(\hat{p}) = \left(1 - \frac{60}{720}\right) \cdot \frac{0,2 \cdot 0,8}{60-1} = \left(1 - \frac{60}{720}\right) \cdot \left(\frac{0,16}{59}\right) =$

$= 0,002485757$ R

$\hat{V}(\hat{A}) = 720^2 \cdot 0,002485757 = 1288,677963$ R

95 % KI för $\hat{A}$: $\hat{A} \pm z_{0,05/2} \cdot \sqrt{\hat{V}(\hat{A})}$

$= 144 \pm 1,96 \cdot \sqrt{1288,677963}$

$= 144 \pm 1,96 \cdot 35,898161$

$= 144 \pm 70,36039556$

$= (73,63960444 : 214,3603956)$ R

$\approx (74 : 214)$

10

b) $N = 600$

Antar $p = 0{,}5$ eftersom det ger maximal varians: $V(p) = 0{,}5 \cdot 0{,}5 = 0{,}25$ BRA!

Minst samma osäkerhet: felmarginalen i a)

Omskrivning från $\hat{A}$ till $\hat{p}$:

$$\frac{70{,}360395\,56}{720} = 0{,}097722777\,16$$

$$1{,}96 \cdot D \leq 0{,}097722777\,16$$

$$D = \frac{0{,}097722777\,16}{1{,}96} = 0{,}049858556\,9 \quad R$$

$$n \geq \frac{600 \cdot 0{,}25}{0{,}049858556\,9^2 \cdot 599 + 0{,}25}$$

$n \geq 87{,}067698\,48$ (lite felmarginal)

$n \geq 88$ Minst 88 hushåll ska väljas

10

20

SU, STATISTIK

Skrivsal: VÄ Anonymkod: 0068-ARH Blad nr: 2

2. Stratifierat urval

a) Eftersom varianserna skiljer sig mellan stratum är optimal allokering att föredra framför proportionell. (Tar alltså hänsyn både till storlek på respektive stratum och dess standardavvikelse.)

$$n_k = n \cdot \frac{N_k \sigma_k}{\sum N_j \sigma_j}$$

$$\sum N_j \sigma_j = 132 \cdot \sqrt{100} + 92 \cdot \sqrt{64} + 27 \cdot \sqrt{25} = 1320 + 736 + 135 = 2191$$

R

$$n_1 = 30 \cdot \frac{1320}{2191} = 18,07393884 \approx 18$$

$$n_2 = 30 \cdot \frac{736}{2191} = 10,07759014 \approx 10$$

$$n_3 = 30 \cdot \frac{135}{2191} = 1,848471018 \approx 2$$

$n_1 + n_2 + n_3 = n$ där $n = 30$

$18 + 10 + 2 = 30$

10

b) Söker: μ $\quad N = 251$

$\bar{y}_1 = 7,2222 \qquad W_1 = \frac{132}{251}$

$\bar{y}_2 = 5 \qquad W_2 = \frac{92}{251}$

$\bar{y}_3 = 2,5 \qquad W_3 = \frac{27}{251}$

$$\bar{y}_{str} = 7,2222 \cdot \frac{132}{251} + 5 \cdot \frac{92}{251} + 2,5 \cdot \frac{27}{251} =$$

$$= 5,899722709$$

$$\hat{V}(\bar{y}_{str}) = \left(\frac{132}{251}\right)^2 \cdot \left(1 - \frac{18}{132}\right) \cdot \left(\frac{36,536}{18}\right) + \left(\frac{92}{251}\right)^2 \cdot \left(1 - \frac{10}{92}\right) \cdot \left(\frac{19,111}{10}\right) +$$

$$+ \left(\frac{27}{251}\right)^2 \cdot \left(1 - \frac{2}{27}\right) \cdot \left(\frac{4,5}{2}\right) = 0,737768944$$

95 % KI: $5,899722709 \pm 1,96 \cdot \sqrt{0,737768944}$

$= 5,899722709 \pm 1,683512212$

$= (4,216210497 : 7,583234921)$

$\approx (4,22 : 7,58)$ /10

c) $\bar{y} = \frac{\sum y_k}{n} = 6,1667$

$\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right) \cdot \frac{s^2}{n} \qquad s^2 = 29,5920$

$\hat{V}(\bar{y}) = \left(1 - \frac{30}{251}\right) \cdot \left(\frac{29,5920}{30}\right) = 0,868503585 7$

95 % KI: $6,1667 \pm 1,96 \cdot \sqrt{0,8685035857}$

$= 6,1667 \pm 1,82659338$ R

$= (4,34010662 : 7,99329338)$ /7

d) Skillnaden mellan felvariansjinalerna är inte så stor eftersom spridningen (varianserna) är stor inom stratum. För att precisionen ska vara bättre vid stratifierat urval än vid OSU ~~förutsätts~~ ju att stratumen har liten varians inom sig (homogena) och stor varians mellan sig (heterogena).
Så är inte riktigt fallet här, stratum 1 har exempelvis nästan exakt lika stor spridning inom sig som hela stickprovet (OSU). BRA 1/3

(30)

3.

a) $\hat{\tau}_{kvot} = \frac{\sum y_k}{\sum x_k} \cdot \tau_x = \frac{182}{130} \cdot 1144 = 1601{,}6$ R

$$\hat{V}(\hat{\tau}_{kvot}) = N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{1}{n} \cdot \left(\frac{\sum (y_k - \hat{R} x_k)^2}{n-1}\right)$$

$\hat{R} = \frac{182}{130} = 1{,}4 \quad N = 104 \quad n = 13$

$$\sum (y_k - \hat{R} x_k)^2 = \sum y_k^2 - 2\hat{R} \sum x_k y_k + \hat{R}^2 \cdot \sum x_k^2$$

$$= 2865 - 2 \cdot \left(\frac{182}{130}\right) \cdot 2020 + \left(\frac{182}{130}\right)^2 \cdot 1444 =$$

$$= 39{,}24$$

$$\hat{V}(\hat{\tau}_{kvot}) = 104^2 \cdot \left(1 - \frac{13}{104}\right) \cdot \left(\frac{1}{13}\right) \cdot \frac{39{,}24}{13-1} = 2380{,}56$$

Standardfelet $= \sqrt{2380{,}56} = 48{,}790982\ldots$ R

10

b) OSU $N=104$ $n=13$

$$\hat{\tau} = N \cdot \bar{y} = 104 \cdot \left(\frac{182}{13}\right) = 1456$$ R

$$\hat{V}(\hat{\tau}) = N^2 \cdot \hat{V}(\bar{y})$$

$$\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$

$$s^2 = \frac{n\sum y_k^2 - \left(\sum y_k\right)^2}{n(n-1)} = \frac{13 \cdot 2865 - 182^2}{13 \cdot (13-1)} =$$

$$= \frac{4121}{156} = 26{,}41666667$$ R

$$\hat{V}(\bar{y}) = \left(1 - \frac{13}{104}\right) \cdot \frac{26{,}41666667}{13} = 1{,}77804487 2$$ R

$$\hat{V}(\hat{\tau}) = 104^2 \cdot 1{,}77804487 2 = 19231{,}33334$$ R

$$\sqrt{\hat{V}(\hat{\tau})} = \text{Standardfelet} = \sqrt{19231{,}33334} =$$

$$= 138{,}67708 3$$ R

OSU ≈ 138,68 som standardfel

Kvot ≈ 48,79 som standardfel

(18)

Ja det lönade sig med kvotskattningen! Eventuell felmarginal kommer bli mycket mindre för kvotskattningen i.o.m. det mycket mindre standardfelet. Väntat eftersom kvotskattning ger bättre precision än OSU om hjälpvariabel x och undersökningsvariabel y har ett starkt samband (som är en rät linje genom origo) vilket detta kan antas vara. /8

Vad händer med punktskattningarna?

4. (samma sak som klusterurval)

a) Gruppurval innebär att man slumpmässigt väljer bland existerande grupper av element man vill undersöka (exempelvis skolor). Vid stratifierat urval delar man in populationen i olika stratan baserat på olika stratifieringsvariabler som kön exempelvis. Sedan drar man slumpmässigt element ur varje strata. Klusterurval kräver ej ram över de enskilda elementen, det gör stratifierat urval. Klusterurval är snabbt och billigt (särskilt om elementen är geografiskt spridda) medan stratifierat urval kan bli krångligt (många stratifieringsvariabler = många strata) och motstridighet mellan stratifieringsvariabler. Klusterurval leder till sämre precision vid skattningar av parametar än OSU), medan stratifierat leder till bättre förutsatt att stratifieringsvariablerna har starka samband med undersökningsvariabeln y. Förutsättningarna: Vid klusterurval ska klustren vara så olika som möjligt inom sig (= hög varians) och så lika som möjligt mellan (= låg varians). För stratifierat urval är det tvärtom: låg varians inom, hög varians mellan. Detta för att öka precisionen i skattningarna. BRA! /5

b) Informera allmänheten om vikten av statistik, informera respondenterna om anonymitet, erbjuda belöningar. Efter: poststratifiering och bortfallsberäkning. Skicka påminnelsebrev under undersökningen. OK /5-

c) CAPI = Computer assisted personal interview
CATI = Computer assisted telephone interview

CAPI är dyrare än CATI, men man har större möjlighet att verifiera att man talar med rätt person. Respondenten känner sig speciell och utvald.

CATI är billigare, mindre chans att veta vem man talar med och under vilka förhållanden (någon misshandlande partner kanske sitter bredvid och lyssnar), respondenten känner sig mindre speciell. Svårt att nå personer via telefon. Dels att så många mobilnummer säkert att folk ogärna svarar på okända nummer. Kan leda till mycket stort bortfall. /5

d) Täckningsfel handlar om kopplingen mellan mål- (problem i) population och rampopulation.
Undertäckning: enheter som finns i målpopulationen finns inte med i urvalsramen. Exempel: vi vill undersöka alla personer som bor i Stockholm. Undertäckning är de som bor här men ej är skrivna i Stockholm. Kan leda till bias (systematiskt fel) i undersökningen. Övertäckning: enheter som inte ingår i målpopulationen finns med i urvalsramen. Samma exempel: övertäckning är exempelvis de som är skrivna i Stockholm men ej bor här. Kan leda till att urvalet blir mindre än man tänkt sig. Oftast inte lika illa som undertäckningens konsekvenser. /5 (Kan exkluderas)

BRA

20

SU, STATISTIK

Skrivsal: VÄ Anonymkod: 0068-ARH Blad nr: 5

5.

a) $\binom{N}{n} = \frac{5!}{2!(5-2)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 3 \cdot 2 \cdot 1} = \frac{5 \cdot 4}{2} = \frac{20}{2} = 10$

10 möjliga stickprov

	Nr	Urval		$\bar{y}$	s^2
AB	1.	2	5	3,5	2,25 + 2,25 = 4,5
AC	2	2	6	4	4 + 4 = 8
AD	3	2	11	6,5	20,25 + 20,25 = 40,5
AE	4	2	16	9	49 + 49 = 98
BC	5	5	6	5,5	0,25 + 0,25 = 0,5
BD	6	5	11	8	9 + 9 = 18
BE	7	5	16	10,5	30,25 + 30,25 = 60,5
CD	8	6	11	8,5	6,25 + 6,25 = 12,5
CE	9	6	16	11	25 + 25 = 50
DE	10	11	16	13,5	6,25 + 6,25 = 12,5

R

BRA!

Exempel på hur jag räknade $\bar{y}$: $\frac{2+5}{2} = 3,5$

b) $\bar{\bar{y}} = \frac{\sum \bar{y}_i}{n} = \frac{80}{10} = 8$ R

$\mu = \sum Y_i \cdot f(Y_i)$

$\mu = 2 \cdot \frac{1}{5} + 5 \cdot \frac{1}{5} + 6 \cdot \frac{1}{5} + 11 \cdot \frac{1}{5} + 16 \cdot \frac{1}{5} = 8$ R

Svar: Ja, $\bar{y}$ är en väntevärdesriktig skattning av μ. I längden blir $\bar{y} = \mu$, alltså 8.

3

3

c)

$$V(\bar{y}) = (\bar{y}_k - E(\bar{y}))^2 \cdot f(\bar{y}_k)$$

(där $E(\bar{y}) = \mu$)

$$V(\bar{Y}) = (3{,}5-8)^2 \cdot \tfrac{1}{10} + (4-8)^2 \cdot \tfrac{1}{10} + (6{,}5-8)^2 \cdot \tfrac{1}{10} + (9-8)^2 \cdot \tfrac{1}{10} + (5{,}5-8)^2 \cdot \tfrac{1}{10} + (8-8)^2 \cdot \tfrac{1}{10} + (10{,}5-8)^2 \cdot \tfrac{1}{10} + (8{,}5-8)^2 \cdot \tfrac{1}{10} + (11-8)^2 \cdot \tfrac{1}{10} + (13{,}5-8)^2 \cdot \tfrac{1}{10} = 9{,}15$$ R

$\sigma_{\bar{y}}^2 / s^2 = \dfrac{\sum (\bar{y}_k - \bar{\bar{y}})^2}{n-1 \;\; N}$ ← N=10 antal möjliga $\bar{y}$

$$(3{,}5-8)^2 + (4-8)^2 + (6{,}5-8)^2 + (9-8)^2 + (5{,}5-8)^2 + (8-8)^2 + (10{,}5-8)^2 + (8{,}5-8)^2 + (11-8)^2 + (13{,}5-8)^2 = 91{,}5$$

$$\frac{91{,}5}{10-1} \approx 10{,}16667$$

Svar: Variansen för den stokastiska variabeln $\bar{Y}$ = 9,15 R

Variansen för stickprovsmedelvärdena = 10,16667

Dessa är inte lika eftersom man vid stickprovsvariansen tar hänsyn till att det är just ett stickprov och delar med n-1 istället för med n.

Stickprovsmedelvärdena uppvisar således, med all rätt, större varians (=större spridning).

2

(8)