STOCKHOLMS UNIVERSITET
Statistiska institutionen
Michael Carlson (examinator)

HT2019

Tentamen i Undersökningsmetodik (4,5 hp)

Kurs: Regressionsanalys och undersökningsmetodik

2020-02-13

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 nyanställd statistiker fick i uppdrag att skatta den genomsnittliga kommunalskatten bland Sveriges $N = 290$ kommuner. Ett obundet slumpmässigt urval utan återläggning av storlek $n = 30$ drogs och kommunalskatten för de dragna kommunerna registrerades och vid en första sammanställning fick man följande resultat:

$$\sum y_k = 661.73 \qquad \sum y_k^2 = 14751.37$$

a) Skatta den genomsnittliga kommunalskatten och ange ett 90 % konfidensintervall. (8p)

b) Med undersökningen och de framräknade resultaten som underlag vill man nu gå ut och meddela att den genomsnittliga kommunalskatt som svenskar betalar är den skattning som du fick i a) ovan (med felmarginal förstås). Vad är det för fel med den slutsatsen? (4p)

c) Anta att man vill ha exakt samma felmarginal som du fick i a) ovan men öka konfidensgraden till 95% istället. Hur stort stickprov skulle du behöva för att få exakt samma felmarginal? Använd att

$$Z_{0.05} \cdot D^* = Z_{0.025} \cdot D$$

där D^* = standardfelet som du fick i a). (8p)

Uppgift 2. (25p)

En läkare vill skatta andelen med för högt blodtryck hos förvärvsarbetande kvinnor i en mindre region. Läkaren gör ett stratifierat urval efter ålder och får följande resultat:

Stratum	Antal i populationen	Urvalsstorlek	Antal med för högt blodtryck
- 25	2510	25	1
26 - 35	5270	75	15
36 - 45	5310	75	39
46 - 55	4960	100	18
56 -	4400	125	36

I dina beräkningar ska ändlighetskorrektion användas.

a) Skatta andelen i hela populationen som har för högt blodtryck och beräkna ett 95% konfidensintervall för andelen. (10p)

b) Anta att läkaren istället hade gjort ett lika stort obundet slumpmässigt urval utan återläggning. Hur stor hade felmarginalen då varit? Blir det bättre eller sämre med stratifieringen? (7p)

c) Läkaren planerar att upprepa undersökningen nästa år. Utgå ifrån de data som du har fått och föreslå läkaren en alternativ stickprovsallokering som kan ge en mindre felmarginal med samma totala stickprovsstorlek. (8p)

Uppgift 3. (25p)

En parkförvaltning vill uppskatta medelåldern på träd i en större stadspark. Att bestämma ålder är besvärligt eftersom man behöver räkna antalet åldersringar som man får genom att ta ett borrprov genom trädet. Dels är det tidskrävande men det kan också vara skadligt för trädet. Men, allmänt gäller att ju äldre trädet, desto större är dess diameter mätt över stammen och diameter är lättare att mäta.

Man mäter diametern på samtliga $N = 1132$ träd i parken och finner att populationsmedelvärdet är $\mu_x = 10.3$. De väljer sedan slumpmässigt $n = 20$ träd för åldersmätning. Låt alltså $x_i =$ diametern mätt i cm och $y_i =$ åldern mätt i år för träd nummer i. Följande summor beräknade från stickprovet finns nu tillgängliga:

$$\sum_{k\in S} x_k = 478 \qquad \sum_{k\in S} y_k = 2\,148$$

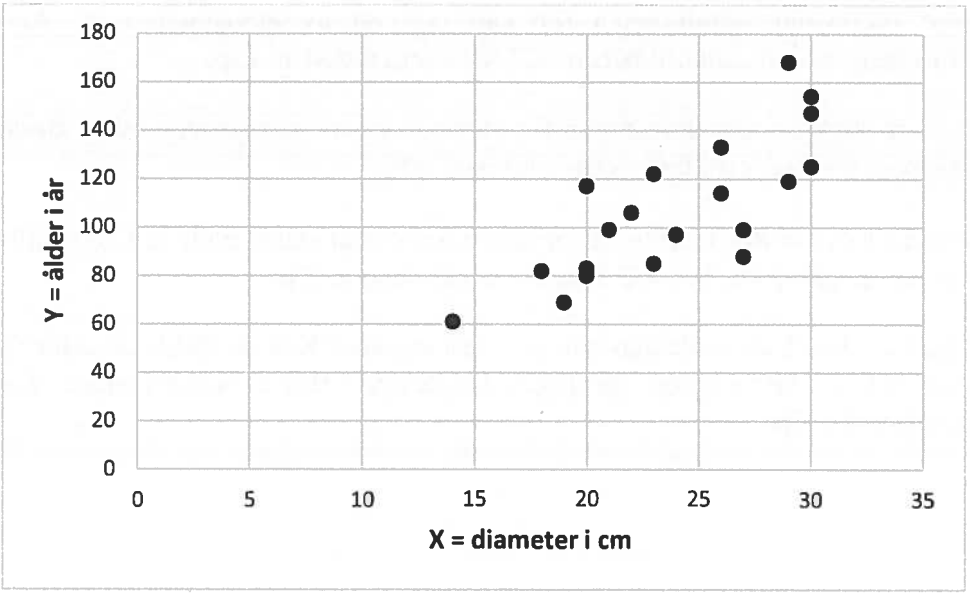
$$\sum_{k\in S} x_k^2 = 11\,832 \qquad \sum_{k\in S} y_k^2 = 246\,304 \qquad \sum_{k\in S} x_k y_k = 53\,315$$

Du ombeds skatta den genomsnittliga åldern för samtliga träd i parken. I beräkningarna ska ändlighetskorrektion användas.

a) Skatta $\mu_y =$ den genomsnittliga åldern med en regressionsestimator med diametern som hjälpvariabel. Beräkna sedan standardfelet för skattningen. (10p)

b) Skatta $\mu_y =$ den genomsnittliga åldern med en kvotestimator med diametern som hjälpvariabel. Beräkna sedan standardfelet för skattningen. (10p)

c) Kommentera dina resultat i a) och b) ovan. Är det stora skillnader? Diskutera kortfattat vilken av de två skattningsmetoderna som det är mest lämplig i denna situation. Utgå ifrån definitionerna av variablerna, de bakomliggande modeller som skattningsmetoderna baseras på och diagrammet nedan som visar stickprovets spridning i bägge variabler. (5p)



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) Förklara vad teleskopeffekten är för något. Vad är bakåt respektive framåt effekter i detta sammanhang? Vad medför det för problem? (5p)

b) Anta att man i en undersökning har en fråga av typen "Vilka av följande alternativ gäller för dig? Fler än ett alternativ kan väljas." Vad ska man tänka på när man definierar de olika svarsalternativen och hur kan svaren kodas? (5p)

c) Ange två olika klassifikationsstandarder som används i svensk statistik och vad de beskriver. Nämn en väsentlig fördel med klassifikationsstandarder. (5p)

d) Beskriv översiktligt vad ett snöbollsurval är och när är en snöbollsdesign kan vara ett användbart alternativ. Är snöbollsurval ett sannolikhetsurval? (5p)

Uppgift 5. (10p)

Man har en population U bestående av $N = 6$ element med följande värden på en variabel Y:

$$U = \{2, 6, 8, 14, 18, 24\}$$

Man föreslår en förenklad urvalsdesign där man endast tillåter tre möjliga stickprov S_k:

$$S_1 = \{2, 14\} \quad S_2 = \{6, 18\} \quad S_3 = \{8, 24\}$$

Vilket stickprov som dras är slumpmässigt och sannolikheten för var och en är lika stor dvs. $1/3$.

a) Ange inklusionssannolikheten för vart och ett av elementen i U. Är den föreslagna urvalsdesignen ett sannolikhetsurval? Motivera ditt svar. (3p)

b) Beräkna stickprovsmedelvärdena för vart och ett av stickproven ovan. Beräkna sedan $\sigma_{\bar{y}}^2 =$ variansen för dessa stickprovsmedelvärden. (2p)

c) Beräkna $V(\bar{y}) =$ den teoretiska variansen för $\bar{y}$ som skulle gälla om du istället hade valt OSU utan återläggning med $n = 2$ som din urvalsdesign. (2p)

d) Vilken av de två urvalsdesignerna ger lägst varians? Kan du förklara varför? TIPS: Vad är den förenklade urvalsdesignen egentligen för design? Hur är listan (ramen) för elementen i U organiserad? (3p)

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	$\hat{\tau} = N\bar{y}$	$V(\hat{\tau}) = N^2 V(\bar{y})$	$\hat{V}(\hat{\tau}) = N^2 \cdot \hat{V}(\bar{y})$
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	$\hat{\tau}_{\text{str}} = N\bar{y}_{\text{str}}$	$\sum_{k=1}^{L} N_k^2 \left(\frac{N_k-n_k}{N_k-1}\right)\frac{\sigma_k^2}{n_k}$	$\sum_{k=1}^{L} N_k^2 \left(1-\frac{n_k}{N_k}\right)\frac{s_k^2}{n_k}$
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.

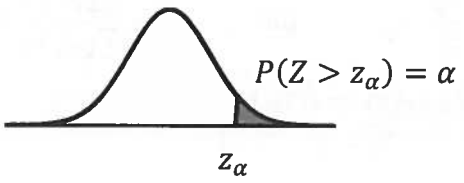
Variansskattning – *formler utgår*

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: 13/2-2020

Sal: Uggleviksalen

Tenta: Undersökningsmetodik

Kurs: Regressionsanalys och undersökningsmetodik

ANONYMKOD: 0016-DUD

☒ 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					7
Lär.ant. 12	21	22	8	4					

POÄNG	BETYG	Lärarens sign.
67	D	MQ

SU, STATISTIK

Skrivsal: Uggleviksalen Anonymkod: 0016-DVD Blad nr: 1

1 $N = 290$ kommuner

$n = 30$ kommuner

$\sum y_k = 661,73$ $\sum y_k^2 = 14751,37$

a) KI = 90% $\Rightarrow t_{\frac{\alpha}{2}} = 1,6449$

$$\bar{y} = \frac{\sum y_k}{n} = \frac{661,73}{30} = 22,057666667 \approx 22,1 \text{ R}$$

$$s^2 = \frac{\sum (y_k - \bar{y})^2}{n-1} = \frac{\sum y_k^2 - n\bar{y}^2}{n-1} = \frac{14751,37 - 30 \cdot 22,057666667^2}{30-1} =$$

$= 5,350008009$ R

$$\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n} = \left(1 - \frac{30}{290}\right)\frac{5,350008009}{30} = 0,159885968 \text{ R}$$

$22,057666667 \pm 1,6449 \cdot 0,159885968$

$= 22,057666667 \pm 0,657124136 =$

$= [21,399942531; 22,715390803] \approx [21,4; 22,7]$ R

Svar: Den genomsnittliga kommunalskatten är 22,1 och ett 90% konfidensintervall ges av $[21,4; 22,7]$

/8

6) Svar: I och med att det endast är en skattning kan man inte vara säker på att det svaret man fick fram gäller för hela Sverige, man har ju endast använt sig av 30 kommuner och det är inte säkert att dessa 30 kommuner speglar hela landet. För att få fram en helt riktig skattning bör man skatta med hjälp av alla 290 kommuner.

I och för sig rätt men inte problemet här /1

$z_{0,05} \cdot D = \pm z_{0,025} \cdot D$

$1,6449 \cdot \sqrt{0,15988529 68} = \pm 0,025 \cdot D$

D i uppgift a) är $\sqrt{0,15988529 68}$

där $D = \sqrt{V(\bar{Y})}$

$\frac{\varepsilon_{max}}{N} = z_{\frac{\alpha}{2}} \sqrt{V(\bar{Y})} \Rightarrow \frac{\varepsilon_{max}}{N} = 0,6577241136$

Se sista bladet

$8 + 1 + 3 = 12$

Sista bladet

SU, STATISTIK

Skrivsal: Aggleviksskolan Anonymkod: 0016-DVD Blad nr: 2

2)

Stratum	Antal i populationen (N_k)	urvalsstorlek (n_k)	Antal m ♪ för högst (klotskyt) (y_k)
-25	2510	25	7
26-35	5270	75	15
36-45	5316	75	39
46-55	4960	100	18
56-	4400	25	36
	22450	400	109

a)

$$\hat{P}_{str} = \sum W_k \cdot \hat{P}_k = \frac{2510}{22450} \cdot \frac{7}{25} + \frac{5270}{22450} \cdot \frac{15}{75} + \frac{5316}{22450} \cdot \frac{39}{75}$$

$$+ \frac{4960}{22450} \cdot \frac{18}{100} + \frac{4400}{22450} \cdot \frac{36}{125} = 0,009472604 +$$

$$+ 0,046948775 + 0,122933185 + 0,059768347$$

$$+ 0,056495435 = 0,270567929 \approx 0,27 \text{ ok}$$

$$\hat{V}(\hat{P}) = \sum W_k^2 \left(1 - \frac{n_k}{N_k}\right) \frac{\hat{P}_k(1-\hat{P}_k)}{n_k - 1} = 0,111804008^2 \left(1 - \frac{25}{2510}\right) \cdot$$

$$\cdot \frac{0,04(1-0,04)}{25-1} + 0,234743875^2 \left(1 - \frac{75}{5270}\right) \cdot \frac{0,2(1-0,2)}{75-1}$$

$$+ 0,236525613^2 \cdot \left(1 - \frac{75}{5310}\right) \cdot \frac{0,52(1-0,52)}{75-1} + 0,220935917^2 \cdot$$

$$\cdot \left(1 - \frac{100}{4960}\right) \cdot \frac{0,18(1-0,18)}{100-1} + 0,194991913^2 \cdot \left(1 - \frac{125}{4400}\right) \cdot \frac{0,288(1-0,$$

$$= 4,585797367 \cdot 10^{-4} \text{ ✓}$$

x_k (N_k/N)	$\hat{p}_k$ ($\frac{1}{n} \cdot Y_k$)
$\frac{2510}{22450} = 0,111804008$	$\frac{1}{25} = 0,04$
$\frac{5270}{22450} = 0,234743875$	$\frac{15}{75} = 0,2$
$\frac{5310}{22450} = 0,236525612$	$\frac{39}{75} = 0,52$
$\frac{4960}{22450} = 0,220935412$	$\frac{18}{100} = 0,18$
$\frac{4400}{22450} = 0,195991091$	$\frac{36}{125} = 0,288$

$kI = 95\% \Rightarrow z_{\frac{\alpha}{2}} = 1,96$

$0,270567929 \pm 1,96\sqrt{4,585797367 \cdot 10^{-4}} =$

$= 0,270567929 \pm 0,041732071 \in [0,2289; 0,3125]$ OK

Svar: Andelen med för högt blodtryck är ca 0,27 och ligger med 95% kI ger $[0,2289; 0,3125]$

10

b) OSU

$$\hat{p} = \frac{1}{n} \cdot \sum y_k = \frac{109}{400} = 0,2725$$ R

$$V(\hat{p}) = \left(\frac{N-n}{N-1}\right) \frac{p(q-p)}{n} = \left(\frac{22450-400}{22450-1}\right) \cdot \frac{0,2725(1-0,2725)}{400}$$

$$= 4,86800622 \cdot 10^{-4}$$ R

= variansen

$$D = \sqrt{4,8600622 \cdot 10^{-4}}$$

För OSU är felmarginalen $1.96 \sqrt{4,8600622 \cdot 10^{-4}}$
och för stratifiering är den
$4585^2 = 3567 \cdot 10^{-4}$, $D^2_{str} < D^2_{osu}$ men nästa

Svar: Det blir något bättre med stratifiering då stratifierings felmarginal är lite mindre.

$\sqrt{\text{Varians}}$ = standardfel = D

felmarginal = $z_{\alpha/2} \cdot D = 1.96 \cdot \sqrt{\text{varians}}$

7

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

$\sigma_k = p(1-p)$

$0,04 \cdot 0,96 = 0,0384$

$0,2 \cdot 0,8 = 0,16$

$0,52 \cdot 0,48 = 0,2496$

$0,18 \cdot 0,82 = 0,1476$

$0,288 \cdot 0,712 = 0,205056$

10+7+4
= 21

$\sum N_j \sigma_j = 2510 \cdot 0,0384 + 5270 \cdot 0,16 + 5310 \cdot 0,2496$
$+ 4960 \cdot 0,1476 + 4400 \cdot 0,205056 = 3087,52$ fel räknat

$n_1 = 400 \cdot \frac{2510 \cdot 0,0384}{3087,52} = 12,48841506 \approx 12$ pers

$n_2 = 400 \cdot \frac{5270 \cdot 0,16}{3087,52} = 109,2397482 \approx 109$ pers

$n_3 = 400 \cdot \frac{5310 \cdot 0,2496}{3087,52} = 171,7075093 \approx 172$ personer

$n_4 = 400 \cdot \frac{4960 \cdot 0,1476}{3087,52} = 94,84583096 \approx 95$ pers

$n_5 = 400 \cdot \frac{4400 \cdot 0,205056}{3087,52} = 116,8894687 \approx 117$ personer

Svar: Den alternativa stratifieringen ger
att $n_1 = 12$ pers $n_2 = 109$ pers $n_3 = 172$ pers
$n_4 = 95$ pers $n_5 = 117$ pers

4

Summa = 505, ej 400

SU, STATISTIK

Skrivsal: VG Anonymkod: 0016-DUP Blad nr: 4

3) $N = 1132$ träd $n = 20$ träd

$\mu_x = 10,3$

x_i = diameter i cm

y_i = åldern mätt i år

$\sum x_k = 478$ $\sum y_k = 2148$ $\bar{x} = 23,9$

$\sum x_k^2 = 11832$ $\sum y_k^2 = 246304$ $\bar{y} = 107,4$

$\sum x_k y_k = 53315$

a) μ_y = genomsnittlig ålder

diameter = hjälpvariabel = x

$\hat{\mu}_{vrg} = \bar{y} + b(\mu_x - \bar{x})$

$$b = \frac{n\sum x_k y_k - (\sum x_k)(\sum y_k)}{n\sum x_k^2 - (\sum x_k)^2} = \frac{20 \cdot 53315 - 478 \cdot 2148}{20 \cdot 11832 - (478)^2} =$$

$$= \frac{39456}{8156} = 4,849926435 \text{ R}$$

$\hat{\mu}_{vrg} = 107,4 + 4,849926435 (10,3 - 23,9) = 41,441000$ R

(≈ 41)

Standardfel:

$$\hat{V}(\hat{\mu}_{vos}) = \left(1 - \frac{n}{N}\right) \cdot \frac{1}{n} \cdot \left(\frac{\sum y_k^2 - n\bar{y}^2 - \hat{b}^2\left(\sum x_k^2 - n\bar{x}^2\right)}{n-2}\right) =$$

$= 16,9175413$ Förklara lite bättre hur du får detta

$\sqrt{16,9175413} = 4,051856525 \approx 4,05$ R

9/10

Svar: Standardfelet är 4,05
och den skattade genomsnittliga
åldern är ca 41 år

b) $\hat{\mu}_{kvot} = \hat{R} \cdot \mu_x = \frac{\sum y_k}{\sum x_k} \cdot \mu_x = \frac{2148}{478} \cdot 37 = 46,7853 = 46$ (≈ 46) R

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

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

$$= 74600 - 2 \cdot \frac{2148}{478} \cdot 53315 + \left(\frac{2148}{478}\right)^2 \cdot 11832$$

$= 6068,357277$ R

(Se nästa sida)

SU, STATISTIK

Skrivsal: VG Anonymkod: 0016-DVD Blad nr: 5

$$\hat{V}(\hat{\mu}_{kvot}) = \left(1 - \frac{20}{1132}\right) \cdot \frac{1}{20} \cdot \frac{6068,3577777}{20-1} =$$

$= 15,68721706$ R

$\sqrt{15,68721706} = 3,960709161 \approx 3,96$ R

Svar: Den genomsnittliga åldern är 46 år och standardfelet är 3,96.

10

c) Svar: Det är inte jätte stor skillnad mellan de olika skattning och inte heller jätte stor skillnad i standardfel. Däremot är kvotskattningens standardfel lite mindre och därför mest lämplig. Som man ser i diagramet är det att utspritt både mellan ålder och mellan längd. Detta ger en hög varians och således ett högt standardfel.

2

10+10+2

22

4) a) Teleskopeffekten är när en resp flyttar en undersökande händelse så att den ska passa in i det tids intervall som är i en undersökning. Det kan t.ex vara en undersökning som handlar om läkarbesök under det senaste tre månaderna. Respondenten kanske då säger att hen haft en läkarbesök de senaste tre månaderna när det egentligen var så män der sedan. Teleskopeffekten bakåt innebär motsatsen. Alltså att en respondent flyttar sin händelse som ägde rum i tidsintervallet så att den inte gör det. Exempel kan vara att respondenten säger att hen ej varit hos läkaren senast 3 månaderna trots att hen varit det. Problemet blir att man får ~~ett intervall~~ en skattning som ej stämmer överens med ~~mål om~~ ~~varm population~~ sant parametervärde (st.prov och pop påverkas ju inte), alltså tecknings problem. man får överteckning vid teleskopeffekten och undertäckning vid teleskopeffekten bakåt.

/4

b) De ska kodas från 0 och uppåt. Alla vet ej skall kodas 9. man ska tänka på att ha så tydlig alternativ som möjligt.

/1 Se lösningsförsl.

SU, STATISTIK

Skrivsal: V6 Anonymkod: 0016-DUP Blad nr: 6

9 c) SSYK-klassifikation av yrken
SSUN-klassifikation av utbildning R

Klassifikationsstandarder gör det väldigt lätt att dela in material i grupper för att sen dra en slutsats. Jämföra resultat i olika undersökn. och i tidsserier /3

d) Snöbollsurval är när man helt slumpmässigt väljer ut en/flera personer. Det kan t ex vara genom att stå på stan. Om frågor om någon vill va med i en undersökning. Det kan vara bra om man vill få in många svarsalternativ. Det är ett sannolikhetsurval där alla människor har teoretiskt samma chans att komma med. Inte nödvändigtvis kända ej korrekt
De är! därför att sannolikheter används

Du beskriver bekvämlighets-/kvoturval /0

8

5) $N = 6$ element

$U = 2, 6, 8, 14, 18, 24$

$S_1 = 2, 14$ $S_2 = 6, 18$ $S_3 = 8, 24$

a) $P(2) = \frac{1}{6}$ $P(6) = \frac{1}{6}$ $P(8) = \frac{1}{6}$ $P(14) = \frac{1}{6}$ $P(18) = \frac{1}{6}$ $P(24) = \frac{1}{6}$ ✓

Svar: Ja, Ja alla urvalen har lika stor chans att bli valda

skall vara större än noll och kända

b)

Urval	$\bar{y}$	$\sigma^2_{\bar{y}}$
2, 14	$\frac{2+14}{2} = 8$	$(16-8)^2 \cdot \frac{1}{3} = 21,333$
6, 18	$\frac{6+18}{2} = 12$	$(24-12)^2 \cdot \frac{1}{3} = 48$
8, 24	$\frac{8+24}{2} = 16$	$(32-16)^2 \cdot \frac{1}{3} = 85,333$

R ?

Svar: $\sigma^2_{\bar{y}} = 21,333$ Stickprovsmedelvärdena $= 8, 12, 16$

48

85,333

SU, STATISTIK

Skrivsal: UG Anonymkod: 0016-DUD Blad nr: 7

c) $V(\bar{Y}) = \left(\frac{N-n}{N-1}\right) \cdot \frac{\sigma^2}{n} = \left(\frac{6-2}{6-1}\right) \cdot \frac{\frac{1}{6}(7-\frac{1}{6})}{2} = 0,055556$ ✓ /2

R R

d) Den som ger lägst varians är OSU. Detta eftersom urvalsdesignen är ett klusterurval Rätt Faktiskt men fel slutsats eftersom fel tidigare och klusterurval har en hög varians inom klustret. Listan för elementen är dessutom redan organiserade Hur? och man skall dra ett urval av kluster. Det är även organiserade så att variansen mellan varannan ska bli så liten som möjligt. /2

4

I/c $z_{0,05} \cdot D^* = z_{0,025} \cdot D$

D = standard felet

$1,6449 \sqrt{0,1594852968} = z_{0,025} \cdot D$

$D^* = 0,1594852968$

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

$\mu_y = \frac{\sum y_k}{N} = \frac{661,73}{290} =$

$2,2818275586$

$\sigma^2 = \frac{\sum y_k^2 - N \cdot \mu_y^2}{N} =$ Fel användning av formeln

$= \frac{1475,37 - 290 \cdot 2,2818275586^2}{290} =$ Använd S_y^2 fr. ej helt enkelt

$45,6600559?$

$n \geq \frac{290 \cdot 45,6600597}{0,1594852968^2 \cdot (290-1) + 45,6600597}$

$n \geq 249,7865164 \approx 250$

Svar: Stickprovet bör vara på 250 kommuner.

orimligt stort?

/3

A