STOCKHOLMS UNIVERSITET VT2020
Statistiska institutionen
Michael Carlson (examinator)

Tentamen i Undersökningsmetodik (4.5 hp)

Kurs: Regressionsanalys och undersökningsmetodik

2020-06-02

Skrivtid:	kl. 15.00 - 21.00 (6 timmar, inklusive en timme för digital överföring)
Godkända hjälpmedel:	Miniräknare, dator, kurslitteratur och föreläsningsanteckningar
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.
  - För full poäng 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.

KONTAKT MED EXAMINATOR UNDER SKRIVNINGEN

- Om det är något som är oklart eller något som verkar konstigt kan du kontakta examinatorn under pågående skrivning via e-post: michael.carlson@stat.su.se eller telefon: 08-16 29 82.
- Undvik att använda Athena för att kontakta examinator.
- Bevaka dock din e-post och Athena under skrivningen för eventuella meddelanden som rör provet.

LYCKA TILL!

Uppgift 1. (25p)

Vid ett visst universitet finns det $N = 20$ studentföreningar. En statistikstudent som läste vid universitetet fick för sig att öva lite undersökningsmetodik genom att studera hur livaktiga studentföreningarna var. Ett stickprov bestående av $n = 4$ föreningar drogs med OSU utan återläggning och följande uppgifter om föreningarnas ekonomiska tillgångar (y) och medlemsantal (m) samlades in:

Förening nummer i	1	2	3	4
y_i = tillgångar i tkr	16.0	96.0	3.6	0.8
m_i = antal medlemmar	116	72	16	24

I dina beräkningar ska ändlighetskorrektion användas.

a) Skatta τ_y = samtliga föreningars totala ekonomiska tillgångar med ett 95% konfidensintervall. (8p)

b) Jämför summan av de ekonomiska tillgångarna i stickprovet med konfidensintervallets nedre gräns som du fick i a) ovan och kommentera. (2p)

c) Hur stort stickprov skulle du behöva för att få en felmarginal för totalen τ_y som är hälften så stor som den du fick i b) uppgiften ovan? Använd stickprovsvariansen för y som en bästa gissning för σ_y^2. (6p)

d) Skatta M = det totala antalet medlemmar i samtliga föreningar med ett 95% konfidensintervall. (8p)

e) Är det rimligt att påstå att urvalet ovan är ett OSU utan återläggning av kluster? TIPS: Hur många kluster kan ett objekt tillhöra? (1p)

Uppgift 2. (25p)

I en kommun genomfördes en opinionsundersökning inför årets kommunalval genom att dra ett urval och fråga varje person om vilket parti de tänker rösta på. Kommunen kan delas in i tre delområden som man historiskt har sett skiljer sig åt politiskt. Områdena kan benämnas Öst (1), Mitten (2) och Väst (3). Populationen av röstberättigade fördelat över de tre områdena är följande:

$$N_1 = 15\,000 \qquad N_2 = 15\,000 \qquad N_3 = 20\,000$$

Man drog ett slumpmässigt urval med $n = 3\,000$. Resultaten från undersökningen ges nedan:

Rösta på parti X	Öst (1)	Mitten (2)	Väst (3)
JA	960	600	120
NEJ	240	600	480

I dina beräkningar ska ändlighetskorrektion användas.

a) Skatta andelen i hela populationen som kommer att rösta på parti X utifrån antagandet att urvalet är ett OSU draget utan återläggning. Beräkna sedan ett 90% konfidensintervall. (8p)

b) Skatta andelen i hela populationen som kommer att rösta på parti X utifrån antagandet att urvalet är ett stratifierat OSU draget utan återläggning. Beräkna sedan ett 90% konfidensintervall. (12p)

c) Du får sedan veta att urvalet faktiskt var draget med vanligt OSU. Förklara hur du ändå kan använda stratifieringen över områden när du skattar andelen i populationen och varför man skulle vilja göra det. Vad blir punktskattningen i så fall? Förklara även hur standardfelet och felmarginalen påverkas jämfört med ett stratifierat urval. Du behöver inte utföra några beräkningar. (5p)

Uppgift 3. (25p)

Efter en stor marknadsföringskampanj för en särskild produkt ville man få preliminära resultat på effekten som kampanjen haft på försäljningen. Man utgick ifrån en population av $N = 60$ försäljningsställen och drog ett OSU utan återläggning av storlek $n = 8$. Man hade uppgift om $x_i =$ försäljningen i tkr under de tre månader som föregick kampanjen och samlade sedan in uppgifter om $y_i =$ försäljningen i tkr under de tre månaderna som följde efter kampanjen. Man fick följande data och någon räknade också ut diverse summor åt dig:

i	1	2	3	4	5	6	7	8
x	183	208	259	273	351	400	440	486
y	514	472	428	363	294	276	239	194

$\sum_{k\in S} x_k = 2\,600$ $\quad$ $\sum_{k\in S} y_k = 2\,780$

$\sum_{k\in S} x_k^2 = 931\,360$ $\quad$ $\sum_{k\in S} y_k^2 = 1\,059\,302$ $\quad$ $\sum_{k\in S} x_k y_k = 815\,227$

Utöver detta visste man från kvartalsredovisningen att den totala försäljningen under de tre föregående månaderna var $\tau_x = 18\,000$ tkr.

I dina beräkningar ska ändlighetskorrektion användas.

a) Skatta $\tau_y =$ den totala försäljningen med en kvotestimator med x som hjälpvariabel. Beräkna sedan standardfelet för skattningen. (12p)

b) Du tycker att resultat i a) är konstigt. Skatta τ_y utan hjälpvariabel, dvs. med ”vanliga” OSU formler. Beräkna sedan standardfelet för skattningen. (8p)

c) Kan du förklara vad skillnaden mellan standardfelen i a) och b) beror på? Är det rätt modell för skattningsförfarandet? Är det möjligt att det blivit något fel någonstans? Om det inte är något fel, kan du föreslå en bättre skattningsmetod? Motivera! (5p)

Uppgift 4. (15p)

För var och en av följande två deluppgifter ska du svara kortfattat. Hela uppgiften bör kunna redovisas på ca en A4-sida, kanske 1½, max 2. Du får gärna komplettera med bilder och skisser.

a) Det finns ett antal s.k. kvalitetskriterier som (enligt lag) ska beaktas och deklareras i Sveriges officiella statistik. Välj ut två av dessa kriterier och beskriv dem kortfattat. Beskriv vilka troliga målkonflikter som kan uppstå mellan de två kriterier som du valt. Om du inte tror att det kan finnas en målkonflikt så måste detta motiveras. (5p)

b) Förklara vad en domän och en domänskattning är och förklara vad som potentiellt kan vara problematiskt när man ska skatta egenskaper i en domän jämfört med situationen med ett stratifierat urval. (5p)

c) Anta att man har ett bortfall på 70% på en fråga om man brukar delta i statistiska undersökningar eller inte (en enkel ja/nej fråga). Resonera lite enkelt kring orsak och verkan beträffande bortfallet för den här frågan. Vidare, anta att 80% av de svarande svarade ’ja’ på frågan. Beräkna minsta och största möjliga andel ’ja’-svar som man teoretiskt kunde ha fått om man inte hade haft något bortfall alls. (5p)

Uppgift 5. (10p)

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

$$U = \{0,1,2,3,4\}$$

a) Lista samtliga stickprov som är möjliga att dra från U med OSU utan återläggning. Beräkna stickprovsvariansen s_y^2 för vart och ett av dessa stickprov. (3p)

TIPS: När stickprovsstorleken är $n = 2$ gäller att $s_y^2 = \frac{(y_i - y_j)^2}{2}$ där y_i och y_j är de två värden som dragits till stickprovet. Eller använd den vanliga variansformeln.

b) Beräkna väntevärdet för s_y^2 genom att beräkna medelvärdet av dina s_y^2 i a) ovan. Beräkna sedan populationsvariansen σ_y^2. Är s_y^2 en väntevärdesriktig skattning av populationsvariansen σ_y^2? Ange en formel för $E(s_y^2)$. (4p)

c) Ange väntevärdet för variansskattningen $\hat{V}(\bar{y})$. (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

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

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

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

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

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 $i = 1,2 \dots, N$
$\bar{m} =$ stickprovsmedelvärde av klusterstorlekarna m_i
$s_{m_i}^2 =$ 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 $i = 1,2 \dots, N$
$\bar{\tau} =$ stickprovsmedelvärdet av totalvärdena τ_i
$s_{\tau_i}^2 =$ 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	$\widehat{M}_{\text{vvr}} = N \cdot \bar{m}$	$\hat{V}(\widehat{M}_{\text{vvr}}) = N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{s_{m_i}^2}{n}$
μ	$\bar{y}_{\text{vvr}} = \frac{\hat{\tau}_{\text{vvr}}}{M} = \frac{N\bar{\tau}}{M}$	$\hat{V}(\bar{y}_{\text{vvr}}) = \frac{N^2}{M^2} \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{s_{\tau_i}^2}{n}$
	$\bar{y}_{\text{kvot}} = \frac{\hat{\tau}_{\text{vvr}}}{\widehat{M}} = \frac{\sum_{i \in S} \tau_i}{\sum_{i \in S} m_i}$	$\hat{V}(\bar{y}_{\text{kvot}}) = \left(\frac{1}{\bar{m}}\right)^2 \left(1 - \frac{n}{N}\right) \frac{\sum_{i \in S} (\tau_i - \bar{y}_{kvot} m_i)^2}{n(n-1)}$
τ	$\hat{\tau}_{\text{vvr}} = N\bar{\tau}$	$\hat{V}(\hat{\tau}_{\text{vvr}}) = N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{s_{\tau_i}^2}{n}$
	$\hat{\tau}_{\text{kvot}} = M\bar{y}_{\text{kvot}} = \frac{M}{\widehat{M}} \hat{\tau}_{vvr}$	$\hat{V}(\hat{\tau}_{\text{kvot}}) = \left(\frac{M}{\bar{m}}\right)^2 \left(1 - \frac{n}{N}\right) \frac{\sum_{i \in S} (\tau_i - \bar{y}_{kvot} m_i)^2}{n(n-1)}$
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 $\hat{R} = \frac{\sum_{k\in S} y_k}{\sum_{k\in S} x_k} = \frac{\hat{\tau}_y}{\hat{\tau}_x}$ $\hat{V}(\hat{\tau}_{\text{kvot}}) = N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{1}{n} \cdot \left(\frac{\sum_{k\in S}\left(y_k - \hat{R}x_k\right)^2}{n-1}\right)$ där $\sum_{k\in S}\left(y_k - \hat{R}x_k\right)^2 = \sum_{k\in S} y_k^2 \; - \; 2\hat{R}\sum_{k\in S} x_k y_k \; + \; \hat{R}^2 \sum_{k\in S} x_k^2$
μ_y	$\hat{\mu}_{\text{kvot}} = \hat{R} \cdot \mu_x = \frac{\sum_{k\in S} y_k}{\sum_{k\in S} x_k} \cdot \mu_x = \frac{\mu_x}{\bar{x}} \cdot \bar{y}$ $\hat{V}(\hat{\mu}_{\text{kvot}}) = \left(1 - \frac{n}{N}\right) \cdot \frac{1}{n} \cdot \left(\frac{\sum_{k\in S}\left(y_k - \hat{R}x_k\right)^2}{n-1}\right)$

Regressionsskattning under OSU u.å.:

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

Poststratifiering under OSU u.å.:

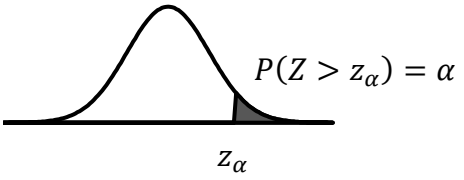
Parametrar och estimatorer - se under **Stratifierat urval** ovan
OBS! Populationsvikterna W_k måste vara kända.
Variansskattning – *formler utgår*

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