STOCKHOLMS UNIVERSITET HT2022
Statistiska institutionen
Michael Carlson (examinator)

Tentamen i Undersökningsmetodik (4,5 hp)

Kurs: Regressionsanalys och undersökningsmetodik

2023-01-13

Skrivtid:	kl. 14.00 - 19.00 (5 timmar)
Godkända hjälpmedel:	Miniräknare utan lagrade formler och lagrad text
Vidhäftade hjälpmedel:	Formelsamling och Statistiska tabeller (endast de tabeller som krävs)

- Tentamen består av 4 uppgifter uppdelade i deluppgifter. Maximalt antal poäng anges per deluppgift.

- Svar med fullständiga redovisningar och motiveringar ska lämnas.
  - Använd endast skrivpapper som tillhandahålls i skrivsalen.
  - Max en uppgift per blad, t.ex. uppgift 1a-d på ett eller flera blad men börja på nytt blad för nästa uppgift 2a-d osv.
  - 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. tabellerna men utgå då ifrån det som är givet. Du kan dock med fördel avrunda det slutliga svaret.

- Maxpoäng är 100 och för godkänt resultat krävs minst 50 poäng. 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. (30p)

Du har i uppdrag att analysera företagen inom en viss bransch och dessas ekonomi. Målpopulationen består av $N = 200$ företag och du ska skatta $\mu_y =$ den genomsnittliga omsättningen för samtliga företag under det senaste året men även $\tau_y =$ den totala omsättningen för samtliga företag i populationen.

Ett OSU u.å. omfattande $n = 30$ företag drogs och följande observationer $y_k =$ omsättning för företag nummer k i miljoner kronor (mkr) erhölls:

3	6	5	4	1	5	3	5	8	4
7	2	6	4	7	9	7	2	10	5
6	8	4	5	6	9	1	5	8	3

$$\sum_{k\in S} y_k = 158 \qquad \sum_{k\in S} y_k^2 = 996$$

a) Beräkna en punktskattning och ge ett 95% konfidensintervall för μ_y. (8p)

b) Beräkna en punktskattning och ge ett 95% konfidensintervall för τ_y. (2p)

Din uppdragsgivare ansåg att resultatet ovan var för oprecist, att det var alldeles för stor osäkerhet.

c) Hur stor stickprovsstorlek krävs för att *felmarginalen* ska vara högst hälften av det du fick i b)-uppgiften ovan? Använd dina resultat från a-b) uppgifterna ovan som din bästa gissning i dina beräkningar. (6p)

OBS! Om du inte har beräknat a-b) ovan, kan du få delpoäng om du förklarar hur du skulle gå till väga för att lösa uppgiften och vilka siffror du skulle använda.

Din uppdragsgivare ville också skatta $P =$ andelen företag i populationen vars omsättning var minst 8 mkr, dvs. andelen företag sådana att $y_k \geq 8$ mkr.

d) Skatta och ge ett 95% konfidensintervall för P. (8p)

I samband med datainsamlingen hade man även frågat efter VD:s kön eftersom man ville skatta genomsnittlig omsättning bland företag med kvinnliga respektive manliga företagsledare. Här hade man utgått från s.k. juridiskt kön som endast omfattar de två kategorierna ”Kvinna” respektive ”Man”.

e) Kategorierna brukar kallas, ja vadå? Vad är problemet med att skatta en parameter för en eller flera kategorier som man inte tagit hänsyn till i urvalsdesignen? (6p)

Uppgift 2. (20p)

Ekonomin för svenska hushåll antas ha påverkats negativt den sista tiden med kraftigt höjda el- och drivmedelspriser och räntehöjningar. Man vill därför analysera hushållens sparande och hur det har förändrats. För att snabbt få fram några siffror har man lyckats definierat en population i ett visst geografiskt område inom en medelstor kommun. Populationen antar man är mer eller mindre representativt för hela landet. Populationen består av $N = 1000$ hushåll.

Man delade in populationen i $L = 2$ strata beroende på $x_k =$ hushållets bruttoinkomst baserat på taxeringsdata från skatteverket. Stratum 1 bestod av de hushåll som hade en bruttoinkomst på högst 350 000 kr per år och Stratum 2 bestod av de hushåll vars bruttoinkomst var högre. I stratum 1 fanns $N_1 = 600$ familjer och i stratum 2 fanns $N_2 = 400$ familjer. Man ansåg sig ha råd med totalt $n = 100$ observationer. Dessa fördelades med hjälp av proportionellt stratifierat urval.

Efter genomförd datainsamling fick man följande stickprovsmedelvärden och stickprovs-standardavvikelser för respektive stratum för $y_k =$ sparande i kronor under december månad 2022 för de tillfrågade hushållen:

h	Gränser	N_h	W_h	n_h	$\bar{y}_h$	s_h
1	$\leq$ 350 tkr	600			90	50
2	$>$ 350 tkr	400			135	70
Summa		1000		100		

Notera att det saknas uppgifter i tabellen ovan, du får fixa dem på egen hand!

a) Beräkna en punktskattning och ge ett 95% konfidensintervall för μ_y, det genomsnittliga sparandet för hela populationen under den aktuella perioden. (10p)

December är ju typiskt en månad där utgifterna ökar och sparandet minskar pga. julen. Man vill därför upprepa undersökningen för januari 2023. Man undrar om man ska allokera annorlunda då.

b) Använd all tillgänglig information och beräkna en ny allokering av stickprovet som är så optimal som möjlig inför nästa undersökning. Stickprovet ska vara lika stort. (5p)

c) Förklara kortfattat när ett stratifierat OSU u.å. är bättre en ett vanligt OSU u.å. Hur skulle du definiera ”bättre” när du ska jämföra urvalsdesigner? (5p)

Uppgift 3. (20p)

Man önskade studera livsmedelsbutikernas ekonomi för ett visst år i en viss kommun. Bland annat var man intresserad av deras personalkostnader. Totalt fanns $N = 90$ butiker och ur denna population drogs ett OSU u.å. av storlek $n = 9$. För dessa utvalda butiker erhölls inte bara uppgift om $y_k =$ personalkostnaden under det förra året utan också om $x_k =$ omsättningen under samma period, allt i miljoner kronor (mkr). Följande tabell sammanställdes:

k	x_k	y_k	x_k^2	y_k^2	$x_k y_k$
1	33	4,6	1089	21,2	151,8
2	126	12,4	15876	153,8	1562,4
3	111	13,2	12321	174,2	1465,2
4	85	9,6	7225	92,2	816
5	23	4,8	529	23,0	110,4
6	58	6,4	3364	41,0	371,2
7	39	4,0	1521	16,0	156
8	11	2,4	121	5,8	26,4
9	61	5,0	3721	25,0	305
Summa	547,0	62,4	45767	552,1	4964,4

Totalt omsatte butikerna $\tau_x = 4\,500$ mkr.

a) Skatta $\tau_y =$ den totala personalkostnaden med en kvotskattning där omsättning används som hjälpvariabel. Beräkna sedan standardfelet för punktskattningen. (10p)

b) Skatta $\tau_y =$ den totala personalkostnaden med en regressionsskattning där omsättning används som hjälpvariabel. Beräkna sedan standardfelet för punktskattningen. (10p)

c) Åskådliggör de nio observationsparen i ett spridningsdiagram, Det behöver inte vara perfekt men det ska inte heller vara överdrivet slarvigt! Argumentera sedan vilken av de två skattningsmetoderna som borde fungera bäst. (5p)

Uppgift 4. (25p)

För var och en av följande deluppgifter ska du svara kortfattat. Hela uppgiften bör kunna redovisas på maximalt ca två A4-sidor. Inga beräkningar behövs men du får gärna illustrera med bilder och skisser om det underlättar, även formler dvs. matematisk framställning går bra om du tycker att det är relevant.

a) Förklara begreppet *systematiskt urval*, hur det går till och varför man ibland använder systematiska urval. Ange en för- och nackdelar det kan ha. (5p)

b) Förklara begreppet *mixed mode*; i vilket sammanhang förekommer det? Ange två fördelar och två nackdelar med en mixed mode design. (5p)

c) *Registerdata* kan ofta användas som ett alternativ till primärdatainsamling. Ange två potentiella fördelar och två nackdelar med att använda registerdata istället för att genomföra en stickprovsundersökning. Register kan även användas som ett hjälpmedel vid statistiska undersökningar. Ge ett exempel. (5p)

d) En känd influencer bad lyssnarna av hennes podcast att via e-post besvara frågan ”Om du kunde, skulle du välja att skaffa barn igen?”. Cirka 200 individer svarade och av dessa svarade ca 70% ”Nej” på frågan. Ca 80% av de svaradande var kvinnor.

   Kommentera (i) upplägget/designen/frågeformuleringen och (ii) resultatet. Var kritisk! Vilka slutsatser, om några, kan man dra från den här undersökningen? (5+5p)

Obs! Om du är osäker på hur mycket som krävs för varje delfråga så reflektera över hur många poäng du kan få för varje delfråga. Du behöver alltså inte skriva långt och mycket, bara det viktigaste!

Formel- och tabellsamling

DESKRIPTIV STATISTIK

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

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:	$\sigma_{xy} = Cov(x,y) = \frac{\sum_{k\in U}(x_k - \mu_x)(y_k - \mu_y)}{n-1} = \frac{\sum_{k\in U} x_k y_k - n\bar{x}\bar{y}}{n-1}$		
	$s_{xy} = Cov(x,y) = \frac{\sum_{k\in U}(x_k - \bar{x})(y_k - \bar{y})}{n-1} = \frac{\sum_{k\in U} x_k y_k - n\bar{x}\bar{y}}{n-1}$		
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 k, ex. $\sum y_k = \sum_{k\in S} y_k$

OBUNDET SLUMPMÄSSIGT URVAL u.å. (HT)

Parameter	Punktskattning	Teoretisk 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: $n \geq \dfrac{N\sigma^2}{D^2(N-1) + \sigma^2}$

STRATIFIERAT URVAL u.å. (HT)

Notation: $L =$ antal strata

$N_h =$ populationsstorleken för stratum $h = 1, \dots, L$

$n_h =$ stickprovets storlek i stratum $h = 1, \dots, L$

$W_h = N_h/N$

$\bar{y}_h =$ stickprovsmedelvärde i stratum $h = 1, \dots, L$

$s_h^2 =$ stickprovsvarians i stratum $h = 1, \dots, L$

Parameter	Punktskattning	Variansskattning $\hat{V}(\cdot)$
μ	$\bar{y}_{\text{str}} = \sum_{h=1}^{L} W_h \bar{y}_h$	$\hat{V}(\bar{y}_{\text{str}}) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}$
τ	$\hat{\tau}_{\text{str}} = N\bar{y}_{\text{str}}$	$\hat{V}(\hat{\tau}_{\text{str}}) = \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}$
P	$\hat{p}_{\text{str}} = \sum_{h=1}^{L} W_h \hat{p}_h$	$\hat{V}(\hat{p}_{\text{str}}) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}$
A	$\hat{A}_{\text{str}} = N\hat{p}_{\text{str}}$	$\hat{V}(\hat{A}_{\text{str}}) = \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}$

Optimal allokering: $n_h = n \cdot \dfrac{N_h\sigma_h}{\sum_{j=1}^{L} N_j\sigma_j}$

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
$s_m^2 =$ 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 $i = 1,2 \dots, N$
$\bar{\tau} =$ stickprovsmedelvärde av totalvärdena τ_i
$s_\tau^2 =$ 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)
vvr står för den vanliga HT-estimatorn under OSU u.å.

Parameter	Punktskattning	Variansskattning
M	$\hat{M}_{\text{vvr}} = N \cdot \bar{m}$	$\hat{V}(\hat{M}_{\text{vvr}}) = N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{s_m^2}{n}$
μ	$\bar{y}_{\text{vvr}} = \frac{\hat{\tau}_{\text{vvr}}}{M} = \frac{N\bar{\tau}}{M}$	$\hat{V}(\bar{y}_{\text{vvr}}) = \frac{N^2}{M^2} \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{s_\tau^2}{n}$
	$\bar{y}_{\text{kvot}} = \frac{\hat{\tau}_{\text{vvr}}}{\hat{M}_{\text{vvr}}} = \frac{\sum_{i \in S} \tau_i}{\sum_{i \in S} m_i}$	$\hat{V}(\bar{y}_{\text{kvot}}) = \left(1 - \frac{n}{N}\right) \cdot \frac{1}{\bar{m}^2} \cdot \frac{\sum_{i \in S} (\tau_i - \bar{y}_{\text{kvot}} m_i)^2}{n(n-1)}$
		där $\sum_{i \in S} (\tau_i - \bar{y}_{\text{kvot}} m_i)^2 =$ [jmfr nästa sida]
		$= \sum_{i \in S} \tau_i^2 \; - \; 2\bar{y}_{\text{kvot}} \sum_{i \in S} \tau_i m_i \; + \; \bar{y}_{\text{kvot}}^2 \sum_{i \in S} m_i^2$
τ	$\hat{\tau}_{\text{vvr}} = N\bar{\tau}$	$\hat{V}(\hat{\tau}_{\text{vvr}}) = N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{s_\tau^2}{n}$
	$\hat{\tau}_{\text{kvot}} = M\bar{y}_{\text{kvot}} = \frac{M}{\hat{M}_{\text{vvr}}} \hat{\tau}_{\text{vvr}}$	$\hat{V}(\hat{\tau}_{\text{kvot}}) = \left(\frac{M}{\bar{m}}\right)^2 \left(1 - \frac{n}{N}\right) \frac{\sum_{i \in S} (\tau_i - \bar{y}_{\text{kvot}} m_i)^2}{n(n-1)}$
P	*formler utgår*	
A	*formler utgår*	

SKATTNINGSMETODER

Notation: τ_y = totalvärdet för variabeln y för hela populationen
$\hat{\tau}_y$ = HT-skattningen av τ_y under OSU
μ_y = populationsmedelvärdet av för variabeln y
Motsvarande beteckningar gäller för variabeln x

Kvotskattning under OSU u.å.:

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

Regressionsskattning under OSU u.å.:

Parameter	Punkt- och variansskattning
μ_y	$\hat{\mu}_{\text{reg}} = \bar{y} + b(\mu_x - \bar{x})$ där $b = \frac{\sum_{k\in S}(y_k - \bar{y})(x_k - \bar{x})}{\sum_{k\in S}(x_k - \bar{x})^2}$ (se sid 7) $\hat{V}\left(\hat{\mu}_{\text{reg}}\right) = \left(1 - \frac{n}{N}\right) \cdot \frac{1}{n} \cdot \left(\frac{\sum_{k\in S}(y_k - \bar{y})^2 - b^2 \sum_{k\in S}(x_k - \bar{x})^2}{n-2}\right)$ där $\sum_{k\in S}(y_k - \bar{y})^2 = \sum_{k\in S} y_k^2 - n\bar{y}^2 = (n-1)s_y^2$
τ_y	$\hat{\tau}_{\text{reg}} = N \cdot \hat{\mu}_{\text{reg}} = N\bar{y} + Nb(\mu_x - \bar{x})$ $\hat{V}\left(\hat{\tau}_{\text{reg}}\right) = N^2 \cdot \hat{V}\left(\hat{\mu}_{\text{reg}}\right)$

Poststratifiering under OSU u.å.:

Parametrar och punktskattning - se under **Stratifierat urval**

OBS! Populationsvikterna W_h måste vara kända.

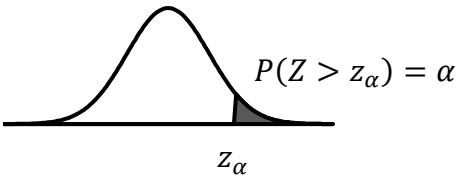
Variansskattning – *formler ingår inte i kursen*

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