Tentamen i Surveymetodik 732G26

Måns Magnusson

11 juni 2014, kl. 8.00-12.00

Surveymetodik med uppsats, 15 hp
Kandidatprogrammet i Statistik och dataanalys
VT2014

Instruktioner

- **Hjälpmedel:**
  - Lohr, S: *Sampling- Design and analysis* (anteckningar får **inte** finnas, men sidflärpar är tillåtet).
  - Miniräknare.
- **Jourhavande lärare:**
  Måns Magnusson
- **Poänggränser:**
  Skrivningen ger maximalt 20 poäng. För betyget godkänt krävs normalt 12 poäng och för betyget väl godkänt krävs 16 p.
- **Övrig information**:
  Samtliga siffror i examen är fiktiva.
  Är det så att någon siffra skulle saknas för att kunna lösa uppgiften, skriv då tydligt ut att du saknar denna information, anta ett godtyckligt värde för denna storhet och lös uppgiften med detta antagande.

Lycka till!


"ERROR: -41"?
THAT'S HELPFUL. IT DOESN'T EVEN SAY WHICH PROGRAM IT'S FROM!
-41? I'LL LOOK IT UP...
IT SAYS -41 IS: "SIT BY A LAKE."
I DON'T KNOW WHERE YOU GOT THAT BOOK, BUT I LIKE IT.
HASN'T BEEN WRONG YET.

Uppgift 1

Aftonbladet är intresserade av att undersöka vilka namn på hundar som är populärast just nu. De planerar därför att göra en webbundersökning där var 1000:e person som går in på Aftonbladet.se ombeds att delta i undersökningen. Datamaterialet kommer således samlas in med en webbenkät.

a) Baserat på förslaget till undersökning. Förklara följande begrepp genom att exemplifiera med studien ovan. Varje begrepp ger **0.5 p**.

i) Bearbetningsfel

ii) Statistikens relevans

iii) Bortfallsfel

iv) Statistikens jämförbarhet

v) Slumpmässigt urval

vi) Domän

vii) Kvotestimation

b) Nämn tre fördelar eller nackdelar med denna design. **1.5p**

Uppgift 2

Riksbanken är intresserad av att försöka förutsäga löneökningstakten i den arbetsföra befolkningen i Sverige (som består av 5428694 personer). De gör detta genom att skicka ut 1000 med obundet slumpmässigt urval och fråga olika personer vad de tror de kommer ha för löneökning nästkommande år (i tusentals kr). Av de som ingick i urvalet svarade 619 personer på enkäten och medelvärdet i urvalet var en löneökning på 0.269 tkr (med en standardavvikelse på 0.251 tkr). Anta "Missing completly at random (MCAR)" och beräkna följande:

a) Beräkna den vad befolkningen tror kommer bli den totala löneökningen nästkommande år med konfidensintervall (99 %). Ignorera bortfallet. **2p**.

b) Beräkna designvikten för respondenterna i denna undersökning. **1p**.

c) Det finns ett intresse av att upprepa undersökningen året efter. Denna gång är de intresserade av en få ett konfidensintervall för $\hat{t}$ på minst $\bar{y}\pm 0.019$. Hur stort antal svarande krävs för att få denna precision. Utgå från resultaten i den undersökning som gjorts sedan tidigare. **2p**

Uppgift 3

Örebro kommun har varit med om ett vattenburet utbrott av calicivirus. De är därför intresserade av att försöka uppskatta hur stor del av befolkningen som har insjuknat under perioden då kommunen vet att utbrottet kan ha ägt rum. Totalt bor det 140599 personer i Örebro.

Smittskyddsenheten i Örebro vill försöka uppskatta hur stor andel av kommuninvånarna (p) som insjuknat och hur många dagar de varit sjuka (y). Totalt deltog $n_r = 876$ personer i undersökningen. Då den troliga smittan är det kommunala vattnet har fokuset lagts på att undersöka de personer som har kommunalt vatten. Undersökningen stratifierades därför baserat på detta. Resultatet av undersökningen framgår nedan.

	N_h	n_h	n_{rh}	$\bar{y}_h$	$s^2{}_h$	p_h
Eget vatten	33420	250	181	0.273	0.064	0.022
Kommunalt vatten	107179	950	695	0.826	0.773	0.285
Samtliga	140599	1200	876	0.712	0.676	0.231

a) Baserat på resultatet ovan beräkna hur stor andel i populationen som insjuknat i calicivirus med tillhörande konfidensintervall (99%). **2p**

b) Beräkna designeffekten för denna skattning. **2p**

c) Hur skulle de ha allokerat urvalet för att skatta $\bar{y}$ i populationen med minsta tänkbara varians? Allokera urvalet (n) till respektive strata med denna allokering? **2p**.

Uppgift 4

Göteborgs kommun är intresserade av att kartlägga hur många mellanstadieelever som är i behov av särskilt stöd från en specialpedagog. För att genomföra denna undersökning har 20 (n) stycken slumpvist utvalda skolor kontaktats och i samtliga skolor har rektor fått frågan om hur många elever som rektor (och anställda) bedömer har behov av specialpedagoger (y).

Totalt finns 210 skolor och 24619 elever i kommunen.

Följande resultat kommer du behöva använda dig av:

$$\begin{aligned} \bar{y}_{\mathcal{S}} &= 22.4 \\ \bar{M}_{\mathcal{S}} = \sum_{i \in \mathcal{S}} \frac{M_i}{n} &= 113.6 \\ s_e^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} e^2 &= 19.2068 \end{aligned}$$

där M_i är antalet elever i skola i och y_i är antalet elever i behov av extra stöd i skola i.

a) Beräkna andelen elever i behov av särskilt stöd med kvotestimatorn. Beräkna även medelfelet för denna skattning. **2.5p**

b) Använd din kvotskattning för att uppskatta det totala antalet elever i behov av speciallärare/specialpedagoger i kommunen. **1.5p**

Lösningar

Uppgift 1 Se föreläsningsanteckningar och kurslitteraturen.

Uppgift 2

a) För att lösa denna uppgift använder vi oss av (2.11 och 2.16) i Lohr [2009, s. 37] för att beräkna variansen. Detta ger:

$$\hat{t} = N\bar{y} = 5428694 \cdot 0.269 = 1460318.686$$

$$\begin{aligned}
\hat{V}(\hat{t}) &= \hat{V}(N \cdot \bar{y}) \\
&= N^2 \cdot \hat{V}(\bar{y}) \\
&= N^2 \cdot \left(1 - \frac{n}{N}\right) \frac{s^2}{n} = \\
&= 5428694^2 \cdot \left(1 - \frac{619}{5428694}\right) \frac{0.251^2}{619} \\
&\approx 54764.48375^2
\end{aligned}$$

Med detta är det sedan möjligt att beräkna konfidensintervallet

$$\begin{aligned}
\hat{t} \pm z_{\alpha/2} \cdot SE(\hat{t}) &= 1460318.686 \pm 2.576 \cdot 54764.48375 \\
&\rightarrow [1319245.37587, 1601391.99613]
\end{aligned}$$

b) För att lösa denna uppgift används resultaten från Lohr [2009, kap. 2.4]. Observera att designvikten beräknas på urvalstorleken (vid designen av studien), inte de faktiskt svarande. Detta ger vid OSU att:

$$w_i = \frac{N}{n} = \frac{5428694}{1000} = 5428.694$$

c) För att lösa denna uppgift använder vi oss av (2.24) och (2.25) i Lohr [2009, s. 47]. Vi är intresserade av att få ett konfidensintervall på 99 % av storleken $\bar{y} \pm$ 0.019 tkr. Detta innebär att $e = 0.019$ i detta fall. Vi behöver också anta standardavvikelse för populationen och här utgår vi från den tidigare undersökningen vilket ger att $S = 0.251$. Detta ger:

$$\begin{aligned}
n_0 &= \left(\frac{z_{\alpha/2} S}{e}\right)^2 \\
&= \frac{2.576^2 \cdot 0.251^2}{0.019^2} \\
&= 1158.06239
\end{aligned}$$

som sedan används för att beräkna det nya n:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

$$\begin{aligned} N & \frac{1158.06239}{1+\frac{1158.06239}{5428694}} \\ = & 1157.8154 \\ \rightarrow & 1158 \end{aligned}$$

Det behövs helt enkelt att 1158 personer **deltar** i studien för att uppnå den efterfrågade precisionen.

Uppgift 3

a) Följande beräkningar kommer vi behöva för att lösa denna uppgift.

	N	$\hat{p}_h$	$\frac{N_h}{N} \cdot \hat{p}_h$	$\left(\frac{N_h}{N}\right)^2$	$1-\frac{n_{rh}}{N_h}$	$\frac{\hat{p}_h(1-\hat{p}_h)}{n_{rh}-1}$	$\left(1-\frac{n_{rh}}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_{rh}-1}$
1	33420.00000	0.02200	0.00523	0.05650	0.99252	0.00009	0.00000
2	107179.00000	0.28500	0.21726	0.58111	0.99114	0.00021	0.00012

Som ett första steg beräknar vi punktskattningen (3.2) i Lohr.

$$\hat{p}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{p}_{str}$$

Detta ger vår punktskattning för andelen insjuknade:

$$\begin{aligned} \hat{p}_{str} & = 0.00523 + 0.21726 \\ & = 0.22249 \end{aligned}$$

Sedan beräknar vi variansen med

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

Hur de olika delarna beräknas framgår i tabellen ovan.

Detta ger således att

$$\begin{aligned} \hat{V}(\hat{p}_{str}) & = 0 + 0.00012 \\ & = 0.00013 \end{aligned}$$

Och konfidensintervallen kan sedan beräknas på följande sätt

$$\begin{aligned} \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\hat{V}(\hat{p}_{str})} & = 0.22249 \pm 2.576 \cdot 0.01134 \\ & \rightarrow [0.19328, 0.25169] \end{aligned}$$

b)

För att beräkna designeffekten används följande från Lohr (7.6) s. 309.

$$deff_\theta = \frac{\hat{V}(\theta)}{\hat{V}_{OSU}(\theta)}$$

för en godtycklig estimator θ.

I vårt fall är $\theta = \hat{p}$. Vi har redan beräknat $\hat{V}(\hat{p}_{str})$ så det som återstår är att beräkna en situation då vi skulle ha ett OSU. Vi använder därför

$$\begin{aligned}\hat{V}_{OSU}(\hat{p}_{str}) &= \left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1} \\ &= \left(1 - \frac{876}{140599}\right)\frac{0.23059(1-0.23059)}{875} \\ &= 0.0002\end{aligned}$$

Nu kan vi enkelt beräkna designeffekten:

$$deff = \frac{\hat{V}_{str}(\hat{p})}{\hat{V}_{OSU}(\hat{p})} = \frac{0.00013}{0.0002} = 0.63707$$

Trots att stratifieringen gjordes "ad hoc" så gav det ändock en tydlig designeffekt.

c) I detta fall finns det ingen skillnad i kostnad mellan de olika urvalen så Neymanallokering är det som bör användas. För att beräkna allokeringen använder vi (3.14) i Lohr s. 89.

$$n_h = \left(\frac{N_h S_h}{\sum_{i=1}^{H} N_i S_i}\right) n$$

Initialt beräknas $\sum_{i=1}^{H} N_i S_i = 102665.601$. Sedan beräknas strata för strata. resultaten framgår av tabellen nedan.

	N_h	${s^2}_h$	$N_h \cdot s_h$	$\left(\frac{N_h \cdot s_h}{\sum N_h \cdot s_h}\right) \cdot n$
1	33420.00	0.06	8455.26	98.83
2	107179.00	0.77	94210.34	1101.17

Således bör urvalet fördelas på följande sätt:

Eget vatten : 99, Kommunalt vatten : 1101

Uppgift 4

a) Vi börjar med att uppskatta kvoten.

$$\hat{B} = \frac{\hat{t}_y}{\hat{t}_x} = \frac{N \cdot \bar{y}_{\mathcal{S}}}{N \cdot \bar{M}_{\mathcal{S}}} = \frac{4704}{23856} = 0.19718$$

Nästa steg är att beräkna variansen med (4.10) i Lohr.

$$\begin{aligned} V(\hat{B}) &= \left(1 - \frac{n}{N}\right) \frac{s_e^2}{\bar{M}_S^2 \cdot n} \\ &= \left(1 - \frac{20}{210}\right) \frac{19.2068}{113.6^2 \cdot 20} \\ &= 0.00007 \end{aligned}$$

vilket ger medelfelet

$$\begin{aligned} SE(\hat{B}) &= \sqrt{V(\hat{B})} \\ &= 0.00821 \end{aligned}$$

b) För att beräkna det totala antalet elever i behov av särskilt stöd använder vi 4.11 i Lohr:

$$\hat{t}_y = \hat{B}M_0 = 0.19718 \cdot 24619 = 4854.4507$$

med variansen

$$\begin{aligned} V(\hat{t}_y) &= V(\hat{B}M_0) \\ &= V(\hat{B}) \cdot M_0^2 \\ &= 0.00007 \cdot 24619^2 \\ &= 40807.82571 \end{aligned}$$

och medelfelet

$$\begin{aligned} SE(\hat{t}_y) &= \sqrt{V(\hat{t}_y)} \\ &= 202.00947 \end{aligned}$$

Appendix

NORMAL CUMULATIVE DISTRIBUTION FUNCTION

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

References

S.L. Lohr. *Sampling: design and analysis*. Thomson, 2 edition, 2009.