Tentamen i Surveymetodik 732G26

Måns Magnusson

15 augusti 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!

Uppgift 1

Fackföreningen Kommunal vill göra en undersökning om långvariga arbetsskador bland sina medlemmar (på grund av tunga lyft m.m.). Totalt har Kommunal 499869 medlemmar och de drar ett slumpmässigt urval av 800 medlemmar och får totalt 637 svar.

Resultaten de fick visade att hela 75% av de svarande hade någon form av arbetsrelaterade besvär. Gör antagandet "Missing completly at random (MCAR)" i undersökningen.

a) Beräkna en skattning av hur stor andel av Kommunals medlemmar som arbetsrelaterade besvär med tillhörande konfidensintervall 90 %. **2p**.

b) Uppskatta det totala antalet medlemmar i Kommunal med arbetsrelaterade besvär med tillhörande konfidensintervall 90 %. **1p**.

g) Det finns ett intresse av att upprepa undersökningen. Kommunal vill göra om undersökningen nästa år och då vill de ha ett konfidensintervall för $\hat{p}$ på minst $\hat{p}\pm 0.021$. Hur stort antal svarande krävs för att få denna precision. Utgå från resultaten i den undersökning som gjorts ovan. **2p**

Uppgift 2

Norrköpings kommun vill undersöka utsatthet för brott i gruppen 20 - 30 år. De har inte bestämt sig för urvaldesign men oroar sig för att de kommer få ett mycket stort bortfall. De tänker sig därför att försöka göra en telefonundersökning.

Baserat på förslaget till undersökning. Förklara följande begrepp genom att exemplifiera med studien ovan.

Observera att det viktigaste är att du visar att du förstår begreppets innebörd och kan sätta det i relation till en undersökning. Varje begrepp ger **0.5 p**.

i) Statistikens relevans

ii) Objektbortfall

iii) Ändlighetskorrektion

iv) Klusterurval

v) Inklutionssannolikhet

iv) Proportionell allokering

Uppgift 3

Statistiska centralbyrån ska genomföra en levnadsnivåundersökning där de vill undersöka hushållsomkostnader för barn i olika åldersgrupper. Det bestämmer sig därför för att

göra ett stratifierat urval av barn i åldrarna 0 - 18 år och drar ett slumpmässigt urval av storlek 4000 och kontaktar föräldrarna genom en postal enkät. Undersökningen är stratifierad efter åldersgrupperna 0 - 6, 7-12 och 13-18 år. Totalt finns det 2066824 barn i denna åldersgrupp och totalt fick Statistiska centralbyrån in $n_r = 2793$ svar.

Undersökningen gav följande resultat:

	N_h	n_h	n_{rh}	$\bar{y}_h$	s_h
0 - 6	806818	2000	1395	2498.79	201.38
7 - 12	641502	1000	696	3566.58	612.13
13 - 18	618504	1000	702	1996.78	1442.04
Samtliga	2066824	4000	2793	2638.70	981.65

a) Baserat på resultatet ovan beräkna ett konfidensintervall (90%) för $\bar{y}_{\mathcal{U}}$. Anta MCAR gällande bortfallet. **2p**

b) Beräkna designvikterna för respektive strata. **1p**

c) Nästa gång undersökningen ska göras så är de intresserade av att få så god precision som möjligt. Vilken allokeringsmetod bör de då använda? Beräkna hur många individer som ska allokeras till respektive strata med denna allokering? **2p**.

Uppgift 4

Linköpings universitet är intresserade av att se hur stora utgifter studenterna har för kursmaterial (datorer, kurslitteratur m.m.) per månad. Då det sedan tidigare är känt att studenter ofta har ett högt bortfall väljer universitet att göra en datainsamling genom att dra ett urval av kurser och samla in data vid föreläsningar.

Totalt finns det 27392 studenter och det ges 913 kurser vid universitetet. För att det ska bli en bra tentauppgift för surveystudenterna i kurs 732G26 väljer universitetet att endast samla in data från 5 kurser. Resultatet från undersökningen kan sammanfattas i nedanstående tabell (där $\bar{y}_i$ är det genomsnittliga utgifterna för respektive kurs och s_i är standardavvikelsen i varje kurs).

	M_i	$\bar{y}_i$	s_i
1	8	259.3	60.7
2	40	335.2	88.6
3	27	309.5	88.6
4	21	381.8	92.5
5	128	334.4	95.3
Samtliga	224	333.3	94.0

a) Skatta hur mycket studenter vid Linköpings universitet lägger på kursmaterial i genomsnitt med tillhörande konfidensintervall (99%). Använd den estimator som är väntevärdesriktig. **2p**

b) Vad har denna skattning för designeffekt? **2p**

c) Gör om skattningen ovan men använd nu kvotestimatorn istället. Beräkna punktskattning, samt tillhörande medelfel. **3p**

Lösningar

Uppgift 1

a) För att lösa denna uppgift använder vi oss av Lohr [2009, s. 37 f.]. Detta ger:

$$\hat{p} = 0.71$$

$$\begin{aligned} \hat{V}(\hat{p}) &= \left(1 - \frac{n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1} = \\ &= \left(1 - \frac{606}{502150}\right) \frac{0.71 \cdot 0.29}{605} \\ &\approx 0.01844^2 \end{aligned}$$

Med detta är det sedan möjligt att beräkna konfidensintervallet

$$\begin{aligned} \hat{p} \pm z_{\alpha/2} \cdot SE(\hat{p}) &= 0.71 \pm 1.96 \cdot 0.01844 \\ &\rightarrow [0.67386, 0.74614] \end{aligned}$$

b) För att lösa denna uppgift använder vi oss också av Lohr [2009, s. 37 f.]. Dock har vi redan beräknat variansen för $\hat{p}$ så den kan vi återanvända. Detta ger:

$$\hat{t} = N\hat{p} = 502150 \cdot 0.71 = 356526.5$$

$$\begin{aligned} \hat{V}(\hat{t}) &= \hat{V}(N \cdot \hat{p}) \\ &= N^2 \cdot \hat{V}(\hat{p}) \\ &= 502150^2 \cdot 0.01844^2 \\ &\approx 9258.09726^2 \end{aligned}$$

Med detta är det sedan möjligt att beräkna konfidensintervallet

$$\begin{aligned} \hat{t} \pm z_{\alpha/2} \cdot SE(\hat{t}) &= 356526.5 \pm 1.96 \cdot 9258.09726 \\ &\rightarrow [338380.62936, 374672.37064] \end{aligned}$$

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 95 % av storleken $\hat{p} \pm 0.027$. Detta innebär att $e = 0.027$ i detta fall. Vi behöver också anta standardavvikelse för populationen och här utgår vi från den tidigare undersökningen vilket ger att $S^2 = (1 - \hat{p}) \cdot \hat{p} = 0.2059$. Detta ger:

$$n_0 = \left(\frac{z_{\alpha/2} S}{e}\right)^2$$

$$
\begin{aligned}
&= \frac{z^2_{\alpha/2}(1-\hat{p})\cdot\hat{p}}{e^2} \\
&= \frac{1.96^2 \cdot 0.2059}{0.027^2} \\
&= 1085.02804
\end{aligned}
$$

som sedan används för att beräkna det nya n:

$$
\begin{aligned}
n &= \frac{n_0}{1+\frac{n_0}{N}} \\
N & \quad \frac{1085.02804}{1+\frac{1085.02804}{502150}} \\
&= 1082.6886 \\
&\rightarrow 1083
\end{aligned}
$$

Det behövs helt enkelt att 1083 personer **deltar** i studien för att uppnå den efterfrågade precisionen.

Uppgift 2

Se föreläsningsanteckningar och kurslitteraturen.

Uppgift 3

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

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

där

	N	$\bar{y}_h$	$\frac{N_h}{N}\cdot\bar{y}_h$	$\left(\frac{N_h}{N}\right)^2$	$1-\frac{n_{rh}}{N_h}$	$\frac{{s_h}^2}{n_{rh}}$	$\left(1-\frac{n_{rh}}{N_h}\right)\left(\frac{N_h}{N}\right)^2\frac{{s_h}^2}{n_{rh}}$
1	806818	2495.71	974.24	0.15	1.00	20.89	3.17
2	641502	3578.21	1110.61	0.10	1.00	353.16	33.97
3	618504	1975.05	591.04	0.09	1.00	2156.55	192.81

Detta ger att:

$$
\begin{aligned}
\hat{\bar{y}}_{str} &= 974.24 + 1110.61 + 591.04 \\
&= 2675.88771
\end{aligned}
$$

Sedan beräknas variansen med hjälp av:

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

Hur de olika delarna beräknas framgår i tabellen.

Detta ger således att

$$\begin{aligned} \hat{V}(\hat{\bar{y}}_{str}) &= 3.17 + 33.97 + 192.81 \\ &= 229.95592 \end{aligned}$$

Och konfidensintervallen kan sedan beräknas på följande sätt

$$\begin{aligned} \hat{\bar{y}}_{str} \pm z_{\alpha/2} \cdot \sqrt{\hat{V}(\hat{\bar{y}}_{str})} &= 2675.88771 \pm 2.576 \cdot 15.1643 \\ &\rightarrow [2636.82448, 2714.95094] \end{aligned}$$

b)

Designvikterna beräknas på följande sätt.

$$d_h = \frac{1}{\pi_h} = \frac{1}{\frac{n_h}{N_h}} = \frac{N_h}{n_h}$$

Det ger följande resultat i vårt exempel:

	N_h	n_h	$\frac{N_h}{n_h}$
1	806818	2000.00	403.41
2	641502	1000.00	641.50
3	618504	1000.00	618.50

c) I detta fall finns det ingen skillnad i kostnad mellan de olika urvalen så Neymanallokering 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 $\sum_{i=1}^{H} N_i S_i = 1454408350.46$. Sedan beräknas strata för strata på följande sätt:

Således bör urvalet fördelas på följande sätt:

0 - 6 : 454, 7 - 12 : 1048, 13 - 18 : 2498

	N_h	s_h	$N_h \cdot s_h$	$\left(\frac{N_h \cdot s_h}{\sum N_h \cdot s_h}\right) \cdot n$
1	806818	204.38	164897462.84	453.51
2	641502	594.27	381225393.54	1048.47
3	618504	1468.52	908285494.08	2498.02

Uppgift 4

a) Vi beräknar först $t_i = m_i \cdot \bar{y}_i$ för alla kluster.

	$\bar{y}_i$	m_i	t_i
1	362.40	38	13771.20
2	342.90	15	5143.50
3	265.20	6	1591.20
4	285.60	66	18849.60
5	277.70	137	38044.90

Vi använder oss av resultaten från Lohr s. 179.

$$\hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{M_0}$$

där M_0 är känd.

$$\begin{aligned} \hat{t}_{unb} &= \frac{N}{n} \sum_{i \in \mathcal{S}} t_i \\ &= \frac{913}{5} (13771.2 + 5143.5 + 1591.2 + 18849.6 + 38044.9) \\ &= 14133313.04 \end{aligned}$$

vilket ger

$$\begin{aligned} \hat{\bar{y}}_{unb} &= \frac{\hat{t}_{unb}}{M_0} \\ &= \frac{14133313.04}{27392} \\ &= 515.96499 \end{aligned}$$

Standardfelet för skattningen är

$$\begin{aligned} V(\hat{t}_{unb}) &= N^2 \left(1 - \frac{n}{N}\right) \cdot \frac{s_t^2}{n} \\ &= 913^2 \cdot \left(1 - \frac{5}{913}\right) \cdot \frac{14345.65188^2}{5} \\ &= 5841354.97211^2 \end{aligned}$$

där

$$\begin{aligned} s_t^2 &= \frac{1}{n-1}\sum_{i\in\mathcal{S}}\left(t_i - \frac{\hat{t}}{N}\right)^2 \\ &= 14345.65188^2 \end{aligned}$$

vilket ger att

$$\begin{aligned} V(\hat{\bar{y}}_{unb}) &= V\left(\frac{\hat{t}_{unb}}{M_0}\right) \\ &= \frac{1}{M_0^2}V\left(\hat{t}_{unb}\right) \\ &= \frac{1}{27392^2}34121427910240.9 \\ &= 213.2504^2 \end{aligned}$$

och ger konfidensintervallet

$$\begin{aligned} \hat{\bar{y}}_{unb} \pm z_{\alpha/2}\cdot\sqrt{\hat{V}(\hat{p}_{str})} &= 515.96499 \pm 1.96\cdot 213.2504 \\ &\rightarrow [97.99421, 933.93578] \end{aligned}$$

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 $\theta = \hat{\bar{y}}_{unb}$. Vi har redan beräknat $\hat{V}(\hat{\bar{y}}_{unb})$ så det som återstår är att beräkna en situation då vi får samma urvalsstorlek med ett OSU. I detta fall innebär det att vi hade haft en urvalsstorlek på $N = 27392$, $n = 262$ och $s = 92.8$.

$$\begin{aligned} \hat{V}_{OSU}(\hat{\bar{y}}) &= \left(1 - \frac{n}{N}\right)\cdot\frac{s^2}{n} \\ &= \left(1 - \frac{262}{27392}\right)\cdot\frac{92.8^2}{262} \\ &= 5.70572^2 \end{aligned}$$

Nu kan vi enkelt beräkna designeffekten:

$$deff = \frac{\hat{V}(\hat{\bar{y}}_{unb})}{\hat{V}_{OSU}(\hat{\bar{y}})} = \frac{45475.73334}{32.55523} = 1396.87968$$

Denna designeffekt säger oss att vi förlorar mycket på att använda klusterurval istället för

OSU i detta fall. Dock kanske kostnaden för att samla in data vid 5 tillfällen väger upp denna osäkerhet tillsammans med att bortfallet blir betydligt mycket lägre.

c) Vi använder här resultaten på s. 180 i Lohr. Observera att nu när vi använder kvotestimatorn måste vi skatta både täljare och nämnare.

$$\begin{aligned}
\hat{\bar{y}}_r &= \frac{\hat{t}_{unb}}{\hat{M}_0} \\
&= \frac{\frac{N}{n}\sum_{i\in\mathcal{S}} \bar{y}_i M_i}{\frac{N}{n}\sum_{i\in\mathcal{S}} M_i} \\
&= \frac{77400.4}{262} \\
&= 295.42137
\end{aligned}$$

Sedan använder vi (5.17) i Lohr, s. 180 för att beräkna variansen för skattningen, vilket ger

$$\begin{aligned}
V(\hat{\bar{y}}_r) &= \left(1-\frac{n}{N}\right)\cdot\frac{1}{n\bar{M}^2(n-1)}\sum_{i\in\mathcal{S}} M_i^2(\bar{y}_i-\hat{\bar{y}}_r) \\
&= \left(1-\frac{5}{913}\right)\cdot\frac{1}{5\cdot 52.4^2\cdot 4}\sum_{i\in\mathcal{S}} M_i^2(\bar{y}_i-\hat{\bar{y}}_r) \\
&= (0.99452)\frac{1}{54915.2}\sum_{i\in\mathcal{S}} M_i^2(\bar{y}_i-\hat{\bar{y}}_r) \\
&= \frac{0.99452}{54915.2}\left(1444\cdot 67^2+225\cdot 47.5^2+36\cdot -30.2^2+4356\cdot -9.79999999999995^2+18769\cdot -17.7^2\right) \\
&= \frac{0.99452}{54915.2}(13321095.94) \\
&= 15.53214^2
\end{aligned}$$

Medelfelet för kvotskattningen är således 15.53214, vilket är nästan en tredjedel så stor som för $\hat{\bar{y}}_{unb}$.

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.