Tentamen i Surveymetodik 732G26

Måns Magnusson

23 mars 2015, kl. 8.00-12.00

Surveymetodik med uppsats, 15 hp
Kandidatprogrammet i Statistik och dataanalys
VT2015

Instruktioner

- **Hjälpmedel:**
  - Lohr, S: *Sampling: Design and analysis.* Anteckningar får **inte** finnas, men små sidflärpar (ett par kvadratcentimeter) med mindre noteringar 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!

Contents

Uppgift 1

Ett undersökningsföretag har fått i uppgift att undersöka hur ofta statsministern syns på bild i de stora dagstidningarna under det senaste året. Totalt har 2520 tidningar publicerats och av dessa görs ett urval på 100 tidningar som undersöks.

Totalt förekommer statsministern i bild i genomsnitt 0.27 tillfällen per tidning (med en standardavvikelse på 0.48938).

a) Beräkna en totalskattning av hur många fotografier på statsministern som presenteras under det senaste året med tillhörande konfidensintervall 90 %. **2p**.

b) Det finns ett intresse av att upprepa undersökningen. Denna gång är de intresserade av en få ett konfidensintervall för $\bar{y}$ på minst $\bar{y}\pm 0.06$. Hur stort antal svarande krävs för att få denna precision. Utgå från resultaten i denna undersökning. **3p**

Lösningsförslag

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:

$$\hat{t} = N\bar{y} = 2520 \cdot 0.27 = 680.4$$

$$\begin{aligned}
\hat{V}(\hat{t}) &= \hat{V}(N \cdot \bar{y}) \\
&= N^2 \cdot \hat{V}(\bar{y}) \\
&= N^2 \cdot \left(1 - \frac{n}{N}\right) \frac{s^2}{n} = \\
&= 2520^2 \cdot \left(1 - \frac{100}{2520}\right) \frac{0.48938^2}{100} \\
&\approx 120.85209^2
\end{aligned}$$

Med detta är det sedan möjligt att beräkna konfidensintervallet

$$\begin{aligned}
\hat{t} \pm z_{\alpha/2} \cdot SE(\hat{t}) &= 680.4 \pm 1.64485 \cdot 120.85209 \\
&\rightarrow [481.61644, 879.18356]
\end{aligned}$$

b) 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 90 % av storleken $\bar{y}\pm$ 0.06. Detta innebär att $e = 0.06$ i detta fall. Vi behöver också anta standardavvikelse för populationen och här utgår vi från den tidigare undersökningen vilket ger att $S = 0.48938$. Detta ger:

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

$$
\begin{aligned}
&= \frac{1.64485^2 \cdot 0.48938^2}{0.06^2} \\
&= 179.98758
\end{aligned}
$$

som sedan används för att beräkna det nya n:

$$
\begin{aligned}
n &= \frac{n_0}{1 + \frac{n_0}{N}} \\
N & \quad \frac{179.98758}{1 + \frac{179.98758}{2520}} \\
&= 167.98918 \\
&\rightarrow 168
\end{aligned}
$$

Det behövs helt ett urval på 168 tidningar för uppnå den efterfrågade precisionen.

Uppgift 2

IOGT-NTO vill genomföra en medlemsundersökning bland sina 30 000 medlemmar. Tanken är att dra ett urval på 500 medlemmar från föreningens register (som innehåller kön och ålder) och sedan genomföra en telefonintervju med de som inkluderats i urvalet. För att förbereda sig har de valt att expertgranska intervjuformuläret och gjort en kognitiv intervju där de bett tre respondenter fylla i intervjuformuläret.

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

- i) Kalibrering
- ii) Kumulativ deltagarandel
- iii) Kvotestimation
- iv) Objektbortfall
- v) Medelfel
- vi) Strata
- vii) Poststratifikation

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

Lösningsförslag

a) och b) Se föreläsningsanteckningar och kurslitteraturen.

Uppgift 3

Linköpings kommun är intresserade av att undersöka hur många som cykelpendlar någon gång i veckan (p) i Linköping och hur många cykelresor längre än 1 km de gjort den senaste veckan (y). Hypotesen är att fler yngre cykelpendlar varför de valt att göra ett stratifierat urval efter ålder av storlek 2000. Totalt deltog $n_r = 1029$ kommuninvånare i undersökningen. Nedan framgår undersökningens resultat. Anta "Missing completely at random" (MCAR) i dina beräkningar.

	N_h	n_h	n_{rh}	$\bar{y}_h$	s_h	p_h
18 - 25	21692	449	160	1.94	1.46	0.86
26 - 40	31389	649	333	0.37	0.60	0.31
41 - 65	43624	902	536	0.16	0.40	0.15
Samtliga	96705	2000	1029	0.50	0.96	0.31

a) Baserat på resultatet ovan beräkna ett konfidensintervall (95%) för $\bar{y}_{\mathcal{U}}$. **2p**

b) Vad kallas den allokering till strata som gjorts i denna undersökning. **1p**

c) Beräkna designeffekten för denna skattning. **1p**

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

Lösningsförslag

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_h	n_{rh}	$\bar{y}_h$	s_h	$\frac{N_h}{N} \cdot \bar{y}_h$	$\left(\frac{N_h}{N}\right)^2$	$1 - \frac{n_{rh}}{N_h}$	$\frac{{s_h}^2}{n_{rh}}$	$\left(1 - \frac{n_{rh}}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{{s_h}^2}{n_{rh}}$
18 - 25	21692	160	1.9	1.5	0.43	0.05	0.99	0.013	0.00066
26 - 40	31389	333	0.37	0.6	0.12	0.11	0.99	0.0011	0.00011
41 - 65	43624	536	0.16	0.4	0.071	0.2	0.99	0.0003	0.000059

Detta ger att:

$$\begin{aligned} \hat{\bar{y}}_{str} &= 0.4346 + 0.11989 + 0.070697 \\ &= 0.62519 \end{aligned}$$

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. Observera att vi använder de svar vi fått (n_{hr}) i respektive strata för att beräkna variansen.

Detta ger således att

$$\begin{aligned}\hat{V}(\hat{\bar{y}}_{str}) &= 0.00066238 + 0.00011085 + 0.000059463 \\ &= 0.00083\end{aligned}$$

Och konfidensintervallen kan sedan beräknas på följande sätt

$$\begin{aligned}\hat{\bar{y}}_{str} \pm z_{\alpha/2} \cdot \sqrt{\hat{V}(\hat{\bar{y}}_{str})} &= 0.62519 \pm 1.96 \cdot 0.02886 \\ &\rightarrow [0.56863, 0.68175]\end{aligned}$$

b) Den allokering som använts är proportionell allokering.

c)

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

$$\begin{aligned}\hat{V}(\hat{\bar{y}}) &= \left(1 - \frac{n_r}{N}\right)\frac{s^2}{n_r} \\ &= \left(1 - \frac{1029}{96705}\right)\frac{0.95586^2}{1029} \\ &= 0.02964^2\end{aligned}$$

Nu kan vi enkelt beräkna designeffekten:

$$deff = \frac{\hat{V}_{str}(\hat{\bar{y}})}{\hat{V}_{OSU}(\hat{\bar{y}})} = \frac{0.02886^2}{0.02964^2} = 0.94789$$

Denna designeffekt säger oss att vi har tjänat på att stratifiera vårt urval (vi får mindre varians vid stratifiering).

d)

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	$\frac{N_h}{n_h}$
18 - 25	21692.00	449	48.31
26 - 40	31389.00	649	48.37
41 - 65	43624.00	902	48.36

Uppgift 4

Skolverket har valt att göra en pilotstudie för att uppskatta antal elever med med läs- och skrivsvårigheter i den svenska skolan.

Totalt finns det 920997 elever i den svenska grundskolan och det finns totalt 4887 grundskolor. Då det endast är en pilotstudie har de valt att endast samla in uppgifter från 4 skolor, men i dessa skolor har de undersökt samtliga elever. Resultatet från undersökningen kan sammanfattas i nedanstående tabell.

	M_i	t_i
1	300	87
2	337	65
3	206	61
4	178	30
Samtliga	1021	243

a) Skatta baserat på denna pilotstudie det totala antalet elever med läs- och skrivsvårigheter. Använd den estimator som är väntevärdesriktig och beräkna tillhörande konfidensintervall (99%). **2p**

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

Lösningsförslag

a) I detta fall är det kluster av samma storlek. För att lösa denna uppgift använder vi oss av (5.1 och 5.3) i Lohr [2009, s. 170 f.] för att beräkna variansen. Detta ger:

$$\begin{aligned}\hat{t} &= \frac{N}{n}\sum_{i}^{N} t_i \\ &= \frac{4887}{4}(87+65+61+30) \\ &= 296885.25\end{aligned}$$

Variansen beräknar vi med (5.3):

$$
\begin{aligned}
Var(\hat{t}) &= N^2 \left(1 - \frac{n}{N}\right) \frac{s_t^2}{n} \\
&= N^2 \left(1 - \frac{n}{N}\right) \frac{\sum_i^n (t_i - \bar{t})^2/(n-1)}{n} \\
&= 4887^2 \left(1 - \frac{4}{4887}\right) \frac{(689.0625 + 18.0625 + 0.0625 + 945.5625)/3}{4} \\
& \quad 57329.41254^2
\end{aligned}
$$

Med detta kan vi sedan beräkna konfidensintervallet.

$$
\begin{aligned}
\hat{t} \pm z_{\alpha/2} \cdot \sqrt{Var(\hat{t})} &= 296885.25 \pm 2.576 \cdot 57329.41254 \\
&\rightarrow [149204.68329, 444565.81671]
\end{aligned}
$$

b) Som ett första steg beräknar vi

$$
\begin{aligned}
\hat{\bar{y}}_r &= \frac{\sum_{\mathcal{S}} t_i}{\sum_{\mathcal{S}} M_i} \\
&= \frac{87 + 65 + 61 + 30}{300 + 337 + 206 + 178} \\
&= 0.238
\end{aligned}
$$

Sedan beräknar vi variansen med (5.17) i Lohr [2009].

$$
\begin{aligned}
Var(\hat{\bar{y}}_r) &= \left(1 - \frac{n}{N}\right) \frac{1}{n\bar{M}^2} \frac{\sum_{\mathcal{S}} (t_i - \hat{\bar{y}}_r M_i)^2}{n-1} \\
&= \left(1 - \frac{4}{4887}\right) \frac{1}{4 \cdot 255.25^2} \frac{243.34 + 231.24 + 143.32 + 152.88}{3} \\
&= 0.03139^2
\end{aligned}
$$

Som ett sista steg kan vi nu beräkna vår total med tillhörande konfidensintervall.

$$
\begin{aligned}
\hat{t} &= \hat{\bar{y}}_r \cdot M_0 \\
&= 0.238 \cdot 920997 \\
&= 219199.09011
\end{aligned}
$$

$$
\begin{aligned}
\hat{t} \pm z_{\alpha/2} \cdot \sqrt{Var(\hat{t})} &= \hat{t} \pm z_{\alpha/2} \cdot \sqrt{Var(M_0 \cdot \hat{\bar{y}}_r)} \\
&= \hat{t} \pm z_{\alpha/2} \cdot \sqrt{M_0^2 \cdot Var(\hat{\bar{y}}_r)} \\
&= 219199.09011 \pm 2.576 \cdot 920997 \cdot 0.03139 \\
&\rightarrow [144736.99879, 293661.18143]
\end{aligned}
$$

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.