Tentamen i Surveymetodik 732G26

Måns Magnusson

Juni 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 företag inom miljöteknik är intresserade av att studera marknadaden för elbilar i Sverige. De väljer att ringa 1200 bilägare från det svenska bilregistret som har 1512301 ägare registrerade. I detta register finns personnummer (alltså kön och ålder) för bilägarna samt de bilar personen äger. I undersökningen visde det sig att 36 respondenter inte hade någon bil (hade nyligen sålt bilen), 414 personer gick inte att nå och 139 ville inte vara med i studien.

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

i) Regressionsestimation

ii) Täckningsfel

iii) Statistikens relevans

iv) Kvoturval

v) Strata

vi) Statistikens samanvändbarhet

vii) Bortfallsfel

viii) Redovisningsgrupp

b) Beräkna hur stor bortfallsandelen är ovan. Utgå ifrån de som inte gått att nå tillhör målpopulationen. **1p**

Lösningsförslag

a) Se föreläsningsanteckningar och kurslitteraturen.

b) Vi använder oss av formeln på s. 10 i Svenska statistikersamfundet. Sektionen för surveystatistik 2005.

$$\begin{aligned} BA &= 1 - \frac{n_s}{n_s + n_b + n_o} \\ &= 1 - \frac{611}{611 + 139 + 414} \\ &= 0.475 \end{aligned}$$

Uppgift 2

TNS-Sifo vill undersöka svenskarnas inställning till reklam. De väljer därför att undersöka hur många i befolkningen (18-70 år - totalt 6563524 personer) som aktivt valt att sätta upp en skylt på dörren med "Ingen reklam, tack". De är också intresserade av att försöka uppskatta hur ofta svenska medborgare upplever att reklam är stötande under en veckas tid.

Undersökningens urvalsstorlek var 2000 och av dessa deltog 937 personer. Av dessa anger 45.464 % att de har en skylt mot reklam på dörren. Respondenterna angav att i genomsnitt upplevde de stötande reklam vid 0.483 tillfällen (med en standardavvikelse på 0.698). I denna undersökning antar vi Missing completly at random (MCAR).

a) Beräkna en totalskattning av hur många i populationen $\hat{t}$ som har en "Ingen reklam, tack!" skylt på sin dörr i populationen med ett konfidensintervall 90 %. **2p**.

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

c) Det finns ett intresse av att upprepa undersökningen nästa år igen. TNS-Sifo vill göra om undersökningen nästa år och då vill de ha ett konfidensintervall för andelen $\hat{p}$ på minst $\hat{p}\pm 0.02$. Hur stort antal svarande krävs för att få denna precision. Utgå från resultaten i denna undersökning när du planerar denna studie. **2p**

Lösningsförslag

a) För att lösa denna uppgift använder vi oss av (2.15, 2.16 och 2.19) i Lohr [2009, s. 37 f.] för att beräkna variansen. Detta ger:

$$\hat{t} = N\hat{p} = 6563524 \cdot 0.455 = 2984056.803$$

$$\begin{aligned}
\hat{V}(\hat{t}) &= \hat{V}(N \cdot \hat{p}) \\
&= N^2 \cdot \hat{V}(\hat{p}) \\
&= N^2 \cdot \left(1 - \frac{n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1} = \\
&= 6563524^2 \cdot \left(1 - \frac{937}{6563524}\right) \frac{0.455 \cdot 0.545}{936} \\
&\approx 106817.845^2
\end{aligned}$$

Med detta är det sedan möjligt att beräkna konfidensintervallet

$$\begin{aligned}
\hat{t} \pm z_{\alpha/2} \cdot SE(\hat{t}) &= 2984056.803 \pm 1.645 \cdot 106817.845 \\
&\rightarrow [2808341.448, 3159772.157]
\end{aligned}$$

b) För att lösa denna uppgift används resultaten från Lohr [2009, kap. 2.4]. Inklutionssannolikheten tar inte "hänsyn" till bortfallet.

$$\pi_i = \frac{n}{N} = \frac{2000}{6563524} = 0.0003047$$

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 90 % av storleken $\hat{p} \pm 0.02$. Detta innebär att $e = 0.02$ i detta fall. Vi behöver också anta standardavvikelse för populationen och här utgår vi från den tidigare undersökningen vilket ger att $S^2 = (1 - \hat{p}) \cdot \hat{p} = 0.248$. Detta ger:

$$\begin{aligned} n_0 &= \left(\frac{z_{\alpha/2} S}{e}\right)^2 \\ &= \frac{z^2_{\alpha/2}(1-\hat{p}) \cdot \hat{p}}{e^2} \\ &= \frac{1.645^2 \cdot 0.248}{0.02^2} \\ &= 1677.348 \end{aligned}$$

som sedan används för att beräkna det nya n:

$$\begin{aligned} n &= \frac{n_0}{1 + \frac{n_0}{N}} \\ N & \quad \frac{1677.348}{1 + \frac{1677.348}{6563524}} \\ &= 1676.919 \\ &\rightarrow 1677 \end{aligned}$$

Det behövs helt enkelt att 1677 personer **deltar** i studien för att uppnå den efterfrågade precisionen.

Uppgift 3

Ett undersökningföretag har fått i uppdrag att undersöka kunder som lämnar ett köpcentrum genom att rekrytera två personer per 200:e som lämnar varuhuset (från två slumpmässigt starttal) och fråga om de kan rekommendera detta varuhus till andra. Totalt så rekryteras på detta sätt 614 personer att ingå i undersökningen. Av dessa är det 415 personer som kommer rekommendera varuhuset.

Utgå från att det totala antalet besökare är jämt delbara med 200.

a) Vad är detta för typ av urvalförfarande? **1p**

b) Utgå från att detta är ett vanligt OSU och skatta hur många (totalen) som kan rekommendera varuhuset med tillhörande konfidensintervall (95 %). **1p**

Du misstänker att det finns periodicitet i hur olika personer lämnar varuhuset och har därför valt att studera resultaten från de två starttalen separat. I båda fall så har du dragit 307 personer och av dessa kan 194 respektive 221 rekommendera varuhuset.

c) Betrakta nu detta som ett klusterurval och skatta hur många (totalen) som kan rekommendera varuhuset med ett tillhörande konfidensintervall (95 %). **2p**

d) Beräkna designeffekten och förklara om du tror att det finns en periodicitet i vilka som lämnar varuhuset. **1p**

Lösningsförslag

a) Systematiskt urval.

b) Vi behöver först beräkna hur stor populationen är vilket vi kan beräkna baserat på n då antalet besökare i varuhuset var jämt delbara med 200.

$$\begin{aligned} N &= n \cdot 100 \\ &= 614 \cdot 100 \\ &= 30700 \end{aligned}$$

Sedan kan vi använda oss av För att lösa denna uppgift använder vi oss av (2.15, 2.16 och 2.19) i Lohr [2009, s. 37 f.] för att beräkna variansen. Detta ger:

$$\hat{t} = N\hat{p} = 30700 \cdot 0.6759 = 20750$$

$$\begin{aligned} \hat{V}(\hat{t}) &= \hat{V}(N \cdot \hat{p}) \\ &= N^2 \cdot \hat{V}(\hat{p}) \\ &= N^2 \cdot \left(1 - \frac{n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1} = \\ &= 30700^2 \cdot \left(1 - \frac{614}{30700}\right) \frac{0.6759 \cdot 0.3241}{613} \\ &\approx 574.51767^2 \end{aligned}$$

Med detta är det sedan möjligt att beräkna konfidensintervallet på följande sätt.

$$\begin{aligned} \hat{t} \pm z_{\alpha/2} \cdot SE(\hat{t}) &= 20750 \pm 1.96 \cdot 574.51767 \\ &\rightarrow [19623.94536, 21876.05464] \end{aligned}$$

c)

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:

$$\begin{aligned}\hat{t} &= \frac{N}{n}\sum_{i}^{N} t_i \\ &= \frac{100}{2}(221+194) \\ &= 20750\end{aligned}$$

Variansen beräknar vi med (5.3):

$$\begin{aligned}Var(\hat{t}) &= N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} \\ &= N^2\left(1-\frac{n}{N}\right)\frac{\sum_i^n (t_i-\bar{t})^2/(n-1)}{n} \\ &= 100^2\left(1-\frac{2}{100}\right)\frac{(182.25+182.25)/1}{2} \\ & 1336.43182^2\end{aligned}$$

d)

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 $\theta = \hat{t}$. Vi har redan beräknat $\hat{V}(\hat{t})$ för klusterurvalet och för situationen med ett vanligt OSU.

$$deff = \frac{\hat{V}(\hat{t})}{\hat{V}_{OSU}(\hat{t})} = \frac{1336.43182^2}{574.51767^2} = 5.41112$$

Då designeffekten är större än 1 kan vi dra slutsatsen att vi har en tydlig periodicitet i vårt systematiska urval.

Uppgift 4

Högskoleverket vill undersöka inträde på arbetsmarknaden efter examen. De genomför därför en undersökning som fått arbete 12 månader efter examen samt den genomsnittliga ingångslönen. Totalt finns det 78036 personer som examinerades för 12 månader sedan, varav 28214 är män och 49822 är kvinnor. Då de vill kunna presentera resultat efter kön har de valt att inkluder 1000 kvinnor och 1000 män i urvalet och av dessa har 303 män och 452 kvinnor svarat. Av

männen har 269 fått arbete inom 12 månader och av kvinnorna har 392 fått arbete inom 12 månader.

a) Baserat på resultatet ovan beräkna ett konfidensintervall (95%) för det totala antalet personer som har fått ett arbete inom 12 månader. **2p**

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

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

Lösningsförslag

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

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

där

	N_h	n_{rh}	$\hat{p}_h$	$\frac{N_h}{N} \cdot \hat{p}_h$	$\left(\frac{N_h}{N}\right)^2$	$1 - \frac{n_{rh}}{N_h}$	$\frac{\hat{p}_h(1-\hat{p}_h)}{n_{rh}-1}$	$\left(1 - \frac{n_{rh}}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_{rh}-1}$
Man	28214	303	0.88779	0.32098	0.13072	0.98926	0.00032987	0.000042657
Kvinna	49822	452	0.86726	0.5537	0.40762	0.99093	0.00025526	0.0001031

Detta ger att:

$$\begin{aligned} \hat{t}_{str} &= N \cdot (0.32098 + 0.5537) \\ &= 68256.53278 \end{aligned}$$

Sedan beräknas variansen med hjälp av:

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

Hur de olika delarna beräknas framgår i tabellen. Observera att vi använder de svar vi fått (n_{hr}) i respektive strata för att beräkna variansen.

Detta ger således att

$$\begin{aligned} \hat{V}(\hat{t}_{str}) &= 78036^2 \cdot 0.000042657 + 0.0001031 \\ &= 942.14^2 \end{aligned}$$

Och konfidensintervallen kan sedan beräknas på följande sätt

$$\begin{aligned}\hat{t}_{str} \pm z_{\alpha/2} \cdot \sqrt{\hat{V}(\hat{t}_{str})} &= 68256.53278 \pm 1.96 \cdot 942.14215 \\ &\rightarrow [66409.93418, 70103.13139]\end{aligned}$$

b)

För att beräkna designeffekten används följande från Lohr (7.6) s. 309. För enkelhetens skull använder jag här $\hat{V}(\hat{p}_{str})$ och $\hat{V}(\hat{p})$.

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

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

$$\begin{aligned}\hat{V}_{OSU}(\hat{p}_{str}) &= \left(1 - \frac{n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1} \\ &= \left(1 - \frac{755}{78036}\right) \frac{0.8755(1-0.8755)}{754} \\ &= 0.01197^2\end{aligned}$$

Nu kan vi enkelt beräkna designeffekten:

$$deff = \frac{\hat{V}_{str}(\hat{p})}{\hat{V}_{OSU}(\hat{p})} = \frac{0.01207^2}{0.01197^2} = 1.01677$$

Denna designeffekt säger oss att i detta fall har vi inte tjänat något på att stratifiera. Vi har till och med fått en något högre designeffekt. Orsaken till detta är att vi inte har en proportionell urvalsdesign.

c)

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	$\frac{N_h}{n_h}$
Man	28214	1000	28.21400
Kvinna	49822	1000	49.82200

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.

Svenska statistikersamfundet. Sektionen för surveystatistik. *Standard för bortfallsberäkning*. Sektionen för surveystatistik, Svenska statiskersamfundet, 2005. URL `http://books.google.se/books?id=OmzQtgAACAAJ`.