Tentamen i Surveymetodik 732G26

Måns Magnusson

12 augusti 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

Folkhälsomyndigheten vill undersöka hur många förskolor som har problem med hårlöss. De är dels intresserade av hur många förskolor som har haft problem med huvudlöss under skolstarten och dels undersöka barnen på förskolan för att se hur många barn de upptäcker med hårlöss.

Totalt finns det 9873 förskolor, myndigheten valde en urvalsstorlek på 500 och av dessa valde 412 förskolor att delta i undersökningen som genomfördes med obundet slumpmässigt urval. Av förskolorna hade 20.388 % haft problem med hårlöss vid skolstarten. När förskolelärarna undersökte huruvida barnen hade hårlöss visade det sig att i genomsnitt hade 1.022 barn hårlöss per förskola (med en standardavvikelse på 1.189). Anta Missing completly at random (MCAR).

a) Beräkna en totalskattning av hur många förskolebarn som haft hårlöss med tillhörande konfidensintervall 95 %. **2p**.

b) Om vi är intresserade av att producera ett konfidensintervall för $\bar{y}$, är urvalsstorleken tillräckligt stor för att konfidensintervallen ska vara approximativt normalfördelade? Varför eller varför inte? **2p**.
För denna uppgift kan följande storhet vara av intresse:

$$\sum_{i \in \mathcal{S}} \frac{(y_i - \bar{y})^3}{n} = 3.192$$

c) Beräkna designvikterna i denna undersökning. **1p**.

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:

$$\hat{t} = N\bar{y} = 9873 \cdot 1.022 = 10090.206$$

$$\begin{aligned}
\hat{V}(\hat{t}) &= \hat{V}(N \cdot \bar{y}) \\
&= N^2 \cdot \hat{V}(\bar{y}) \\
&= N^2 \cdot \left(1 - \frac{n}{N}\right) \frac{s^2}{n} = \\
&= 9873^2 \cdot \left(1 - \frac{412}{9873}\right) \frac{1.189^2}{412} \\
&\approx 566.143^2
\end{aligned}$$

Med detta är det sedan möjligt att beräkna konfidensintervallet

$$\begin{aligned}
\hat{t} \pm z_{\alpha/2} \cdot SE(\hat{t}) &= 10090.206 \pm 1.96 \cdot 566.143 \\
&\rightarrow [8980.565, 11199.847]
\end{aligned}$$

b) För att lösa denna uppgift använder vi oss av (2.23) i Lohr [2009, s. 44] för att på detta sätt uppskatta den minimala urvalsstorleken för när en normalapproximation rekommenderas. Vår urvalsstorlek är 412 och minimum för normalapproximation krävs:

$$\begin{aligned} n_{min} &= 28 + 25\left(\frac{1}{s^3}\sum_{i\in\mathcal{S}}\frac{(y_i-\bar{y})^3}{n}\right)^2 \\ &= 28 + 25\left(\frac{1}{1.189^3}3.192\right)^2 \\ &= 28 + 25\left(1.899\right)^2 \\ &= 28 + 90.152 \\ &\approx 118 \end{aligned}$$

c) För att lösa denna uppgift används resultaten från Lohr [2009, kap. 2.4]. Observera att designvikten beräknas på urvalstorleken, inte de faktiskt svarande. Att utgå från de som svarar innebär att vi antar MCAR. Detta ger vid OSU att:

$$w_i = \frac{N}{n} = \frac{9873}{500} = 19.746$$

Uppgift 2

Ett undersökningsföretag har skapat en ny webbpanel för att kunna genomföra korta, snabba marknadsundersökningar per e-post. De har rekryterat deltagare till panelen genom ett slumpmässigt urval.

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

- i) Kumulativ deltagarandel
- ii) Designvikt
- iii) Poststratifikation
- iv) Slumpmässigt urval
- v) Dominans
- vi) Statistikens relevans
- vii) Ramfel

b) Nämn tre sätt det skulle vara möjligt att förbygga bortfall i denna undersökning. **1.5p**

Lösningsförslag

a) och b) Se föreläsningsanteckningar och kurslitteraturen.

Uppgift 3

Journalistförbundet är intresserade av att undersöka hur stor andel av deras medlemmar som har varit utsatta för hot i samband med deras yrkesutövning. Journalistförbundet delar in sina medlemmar i de som arbetar i dagspress, public service-företagen och i tidskrifter. De och skickar därför ut en postenkät till ett stratifierat slumpmässigt urval (efter i yrkesgrupp) av storlek 1500 av de totala antalet 11181 medlemmar. De har valt en urvalsstorlek på 500 i respektive strata. Totalt deltog $n_r = 790$ medlemmar i undersökningen.

Undersökningen gav följande resultat:

	N_h	n_h	n_{rh}	p_h
Dagspress	5312	500	260	0.44
Public service	1978	500	271	0.27
Tidskrifter	3891	500	259	0.61
Samtliga	11181	1500	790	0.44

a) Baserat på resultatet ovan beräkna ett konfidensintervall (99%) för $\hat{p}_{\mathcal{U}}$. **2p**

b) Vad kallas den allokering de valt? **1p**

De vill nu genomföra en ny undersökning och konsulterar dig. I budgeten har de lagt in att de kan skicka ut totalt 1500 enkäter.

h) De vill att den nya undersökningens urval ska ha samma fördelning mellan strata som fördelningen ser ut i målpopulationen. Vilken allokering föreslår du och hur många respondenter ska allokeras till respektive strata? **2p**.

Lösningsförslag

a) Som ett första steg beräknar vi punktskattningen (3.2) i Lohr [2009].

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

där

Detta ger att:

$$\begin{aligned} \hat{p}_{str} &= 0.21 + 0.0471 + 0.212 \\ &= 0.469 \end{aligned}$$

	N_h	n_{rh}	$\hat{p}_h$	$\frac{N_h}{N} \cdot \hat{p}_h$	$\left(\frac{N_h}{N}\right)^2$	$1 - \frac{n_{rh}}{N_h}$	$\frac{\hat{p}_h(1-\hat{p}_h)}{n_{rh}-1}$	$\left(1 - \frac{n_{rh}}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_{rh}-1}$
Dagspress	5312	260	0.44	0.21	0.23	0.95	0.00095	0.0002
Public service	1978	271	0.27	0.047	0.031	0.86	0.00072	0.00002
Tidskrifter	3891	259	0.61	0.21	0.12	0.93	0.00092	0.0001

Sedan beräknas variansen med hjälp av:

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

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

$$\begin{aligned} \hat{V}(\hat{p}_{str}) &= 0.000204 + 0.0000195 + 0.000104 \\ &= 0.018^2 \end{aligned}$$

Och konfidensintervallen kan sedan beräknas på följande sätt

$$\begin{aligned} \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\hat{V}(\hat{p}_{str})} &= 0.469 \pm 2.58 \cdot 0.018 \\ &\rightarrow [0.423, 0.516] \end{aligned}$$

b) Den allokering de använt är lika allokering.

c) I detta fall bör proportionell allokering användas. För att beräkna allokeringen använder vi:

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

	N_h	$\left(\frac{N_h}{\sum N_h}\right) \cdot n$
Dagspress	5312	713
Public service	1978	265
Tidskrifter	3891	522

Således bör urvalet fördelas på följande sätt:

	n_h
Dagspress	712
Public service	265
Tidskrifter	522

Uppgift 4

Länsstyrelsen i Stockholm har fått regeringens uppdrag att undersöka prostitutionens omfattning i Sverige. Som en del i detta arbete har de valt att med en befolkningsundersökning undersöka antalet sexköpare i den totala populationen. Det är känt sedan tidigare att den absoluta majoriteten av de som köper sex är män varför det är av intresse att uppskatta det totala antalet män som köpt sex.

Ett urval görs från Sveriges befolkning 18 - 65 år (5966951 personer, varav 3032524 och 2934427 kvinnor). Länsstyrelsen väljer att dra ett urval på 4000 och av dessa är det bara 1143 som deltar i undersökningen på grund av frågornas känsliga natur. Av dessa är 553 män. Totalt uppger 7 män att de köpt sex de senaste 12 månaderna. Observera att undersökningen **INTE** är stratifierad på kön.

a) Beräkna det totala antalet män som köpt sex de senaste 12 månaderna med tillhörande konfidensintervall 90 %. **3p**.

b) Utgå nu från att det totala antalet män i populationen är okänd. Beräkna under dessa förhållanden det totala antalet män som köpt sex de senaste 12 månaderna med tillhörande konfidensintervall 90 %. **2p**.

Lösningsförslag

a) I detta fall har vi ett tydligt exempel på en skattning i en redovisningsgrupp. I detta fall använder vi därför den domänestimationsmetod som bygger på kvotestimatorn och som i (4.13) i Lohr [2009] och antar att $n_d/(n_d - 1) \approx 1$. På s. 38 i Lohr [2009] framgår hur s^2 kan beräknas för proportioner vilket vi kan använda oss av. Detta ger att

$$\begin{aligned} \hat{p} &= \frac{\sum_{\mathcal{S}_d} y_i}{n_d} \\ &= \frac{7}{553} \\ &= 0.013 \end{aligned}$$

vilket vi sedan kan multiplicera med det totala antalet män i populationen. Vår skattning är således

$$\begin{aligned} \hat{t}_d &= N_d \cdot \hat{p}_d \\ &= 38386.38 \end{aligned}$$

För att beräkna variansen använder vi (4.13) i Lohr [2009] men för totalen istället.

$$\begin{aligned} Var(\hat{t}_d) &= Var(N_d \cdot \hat{p}_d) \\ &= N_d^2 \cdot Var(\hat{p}_d) \end{aligned}$$

$$
\begin{aligned}
&= N_d^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{\hat{p}_d(1-\hat{p}_d)}{n_d - 1} \\
&= 3032524^2 \cdot \left(1 - \frac{1143}{5966951}\right) \cdot \frac{0.013(1-0.013)}{552} \\
&= 14428.239^2
\end{aligned}
$$

Detta ger följande konfidensinterval:

$$
\begin{aligned}
\hat{t} \pm z_{\alpha/2} \cdot SE(\hat{t}) &= 38386.38 \pm 1.645 \cdot 14428.239 \\
&\rightarrow [14651.927, 62120.832]
\end{aligned}
$$

b) Känner vi inte till det totala antalet män i populationen använder vi s. 134 f. i Lohr [2009]. Detta ger att

$$
\begin{aligned}
\hat{t}_u &= N \cdot \hat{p}_u \\
&= 36543.007
\end{aligned}
$$

och på liknande sätt beräknar vi variansen på samma sätt som vid ett vanligt OSU

$$
\begin{aligned}
Var(\hat{t}_u) &= Var(N \cdot \hat{p}_u) \\
&= N^2 \cdot Var(\hat{p}_u) \\
&= N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{\hat{p}_u(1-\hat{p}_u)}{n-1} \\
&= 5966951^2 \cdot \left(1 - \frac{1143}{5966951}\right) \cdot \frac{0.006(1-0.006)}{1142} \\
&= 13774.308^2
\end{aligned}
$$

Detta ger följande konfidensinterval:

$$
\begin{aligned}
\hat{t} \pm z_{\alpha/2} \cdot SE(\hat{t}) &= 36543.007 \pm 1.645 \cdot 14428.239 \\
&\rightarrow [13884.271, 59201.743]
\end{aligned}
$$

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.