Tentamen i Surveymetodik 732G26

Måns Magnusson

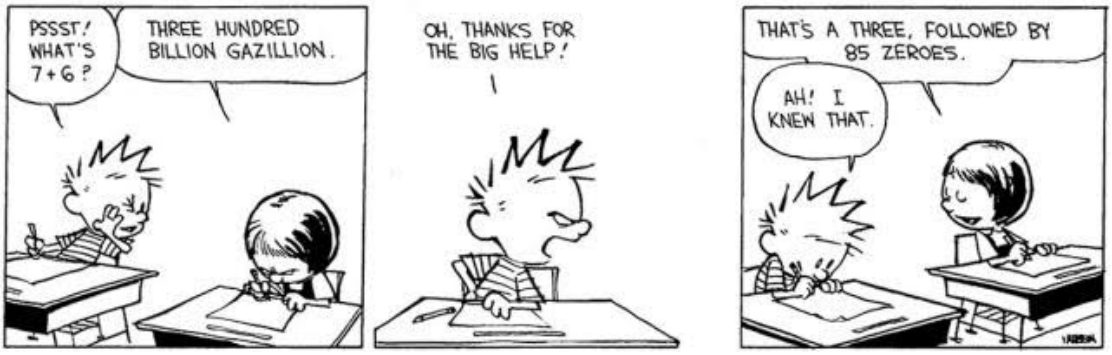
5 juni 2013, kl. 8.00-12.00

Surveymetodik med uppsats, 15 hp
Kandidatprogrammet i Statistik och dataanalys
VT2013

Instruktioner

- **Hjälpmedel:**
  - Lohr, S: *Sampling- Design and analysis* (anteckningar får finnas).
  - Miniräknare.
- **Jourhavande lärare:**
  Tommy Schyman
- **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.

Lycka till!



1. Ett försäkringsbolag är intresserade av att undersöka de samhälleliga kostnader för olyckor på svenska vägar. En av faktorerna de är intresserade är hur många olyckor som resulterat i att en eller flera personer fortfarande har men från olyckan 12 månader efter det att olyckan inträffade. Totalt har 16119 olyckor inträffat och från dessa väljs 200 olyckor ut slumpmässigt för att studeras vidare. I undersökningen framgår att i 32 olyckor har personer fortfarande men 12 månader efter olyckan.

   (a) Beräkna en skattning av andelen olyckor där minst en person fortfarande har men 12 månader efter olyckstillfället med tillhörande konfidensintervall (95 %). **2p**.
   Svar: För att lösa denna uppgift använder vi oss av (2.19) i Lohr [2009, s. 38] för att beräkna variansen. Detta ger:

$$\hat{p} = 0.16$$

$$\hat{V}(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1} = \left(1 - \frac{200}{16119}\right)\frac{0.16 \cdot 0.84}{199} \approx 0.026^2$$

$$\hat{p} \pm z_{\alpha/2}SE(\hat{p}) = 0.16 \pm 1.96 \cdot 0.026 \rightarrow [0.109, 0.211]$$

   (b) Beräkna en skattning av det totala antalet olyckor med tillhörande konfidensintervall (95 %) där minst en person fortfarande har men 12 månader efter olyckstillfället. **1p**.
   Svar: För att lösa denna uppgift använder vi att totalen är proportionen multiplicerat med populationstotalen:

$$\hat{t} = N \cdot \hat{p} = 16119 \cdot 0.16 = 2579.04$$

   På liknande sätt beräknar vi variansen och konfidensintervallen:

$$V(\hat{t}) = V(\hat{p} \cdot N) = N^2 V(\hat{p}) = 16119^2 \cdot 0.026^2 = 419.094^2$$

$$\hat{t} \pm z_{\alpha/2}SE(\hat{t}) = 2579.04 \pm 1.96 \cdot 419.094 \rightarrow [1757.616, 3400.464]$$

   (c) Baserat på denna undersökning vill försäkringsbolaget löpande börja studera denna aspekt av olyckorna. För att kunna identifiera skillnader över tid har de bestämt sig för att konfidensintervallet (95 %) för skattningen av andelen maximalt får vara $\pm$ 0.05, oavsett andelen olyckor i urvalet. Baserat på resultaten från den första undersökningen, beräkna vilken urvalsstorlek som skulle krävas för att uppnå denna precision. **2p**.
   Svar: 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 av storleken $\hat{p} \pm$ 0.05. Detta innebär att $e = 0.05$ i detta fall. Vi behöver också anta standardavvikelse för populationen. Då $p = 0.5$ uppnås det maximala värdet för $S^2 = 0.25$ $(p(1-p))$, för att vara säkra sätter vi därför S^2 till 0.25. Detta ger:

$$n_0 = \left(\frac{z_{\alpha/2}S}{e}\right)^2 = \frac{1.96^2 \cdot 0.25}{0.05^2} = 384.16$$

som sedan används för att beräkna det nya n:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{384.16}{1 + \frac{384.16}{16119}} = 375.218 \rightarrow 376$$

2. Ett marknadsundersökningsföretag är intresserade av att undersöka kostnader i samband med bröllop. Tanken är att genomföra studien som en enkätstudie och samla in data från 300 bröllop (genom att kontakta bröllopsparen). Rampopulationen består av samtliga bröllop i Sverige 2012 vilket var 50123 stycken. Sedan innan är det känt att religiösa bröllop tenderar att vara dyrare än borgerliga vigslar, varför företaget har beslutat att stratifiera efter religiös repsektive borgerlig vigsel. Av de det totala antalet vigslar är 19101 vigslar borgerliga (strata 1) och de övriga 31022 (strata 2) vigslarna religiösa.
Undersökningen har gjort en proportionell allokering av urvalet och resultatet blev $\bar{y}_1 = 164913$, $s_1 = 8419$, $\bar{y}_2 = 192236$ och $s_2 = 14389$. För hela urvalet (då båda strata lagts ihop) blev resultatet $\bar{y} = 175296$, $s = 17279$.

   (a) Beräkna en skattning av det genomsnittliga kostnaden för ett bröllop inklusive ett konfidensintervall (95 %). **2.5p**.
   Svar: Först behöver vi beräkna hur stor andel av urvalet som gjorts till varje strata. För detta används ??? i Lohr:

$$n_1 = n \cdot (N_1/N) = 300 \cdot (19101/50123) \approx 114$$

$$n_2 = n \cdot (N_2/N) = 300 \cdot (31022/50123) \approx 186$$

   Baserat på dessa urvalsstorlekar är det sedan möjligt att beräkna punktskattningen och variansen för skattningen av den genomsnittliga kostnaden för ett bröllop i Sverige:

$$\hat{\bar{y}}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \frac{19101}{50123} \cdot 164913 + \frac{31022}{50123} \cdot 192236 = 181824$$

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

$$\left(1 - \frac{114}{19101}\right) \left(\frac{19101}{50123}\right)^2 \frac{8419^2}{114} + \left(1 - \frac{186}{31022}\right) \left(\frac{31022}{50123}\right)^2 \frac{14389^2}{186} =$$

$$716.655^2$$

   Med punktskattningen och variansen går det att konstruera ett konfidensintervall på följande sätt:

$$\hat{\bar{y}}_{str} \pm z_{\alpha/2} SE(\hat{\bar{y}}_{str}) = 181823.682 \pm 1.96 \cdot 716.655 \rightarrow [180419.038, 183228.325]$$

(b) Beräkna designeffekten för denna studie. **1.5p**.

Svar: På följande sätt beräknas designeffekten (se XXX i Lohr):

$$deff_{\hat{\bar{y}}_{str}} = \frac{\hat{V}(\hat{\bar{y}}_{str})}{\hat{V}(\hat{\bar{y}})}$$

Vi har $V(\hat{\bar{y}}_{str})$ från uppgiften ovan. Vi behöver därför beräkna $V(\hat{\bar{y}})$ vilket vi gör med följande formel:

$$\hat{V}(\hat{\bar{y}}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n} = \left(1 - \frac{300}{50123}\right)\frac{17279^2}{300} = 994.631^2$$

Sätter vi in det i formeln ovan får vi följande resultat:

$$deff_{\hat{\bar{y}}_{str}} = \frac{\hat{V}(\hat{\bar{y}}_{str})}{\hat{V}(\hat{\bar{y}})} = \frac{716.655^2}{994.631^2} = 0.519$$

(c) Beräkna designvikterna i de olika strata. **1p**.

Designvikten är inversen av inkutionssannolikheterna för respektive strata vilket ger följande vikter.

$$w_1 = N_1/n_1 = (19101/114) = 167.553$$

$$w_2 = N_2/n_2 = (31022/186) = 166.785$$

Eftersom det är ett proportionellt urval (ofta kallat självvägande) så blir vikterna nästan identiska mellan strata.

3. Från en (mycket) liten population på bestående av $y = (3,4,1,8,12)$ väljs ett slumpmässigt urval på $n = 3$ utan återläggning. Beräkna/lista de olika urvalen, urvalssannolikheten för respektive urval, urvalfördelningen för $\bar{y}$, det (teoretiska) förväntade värdet för $\bar{y}$ ($E(\bar{y})$) och (den teoretiska) variansen för $\bar{y}$ ($Var(\bar{y})$). **5p**.

Svar: Nedan framgår de första tre delarna i uppgiften.

	Obs 1	Obs 2	Obs 3	$P(\mathcal{S})$	$\bar{y}_{\mathcal{S}}$
1	3	4	1	0.10	2.67
2	3	4	8	0.10	5.00
3	3	4	12	0.10	6.33
4	3	1	8	0.10	4.00
5	3	1	12	0.10	5.33
6	3	8	12	0.10	7.67
7	4	1	8	0.10	4.33
8	4	1	12	0.10	5.67
9	4	8	12	0.10	8.00
10	1	8	12	0.10	7.00

För att beräkna det förväntade värdet används XX i Lohr:

$$E(\bar{y}) = \sum_{\mathcal{S}} P(\mathcal{S}) \cdot \bar{y}_{\mathcal{S}} = 5.6$$

På liknande sätt beräknas den teoretiska variansen (se XX i Lohr):

$$E(\bar{y}) = \sum_{\mathcal{S}} P(\mathcal{S}) \cdot (\bar{y}_{\mathcal{S}} - E(\bar{y}))^2 = 2.573$$

4. Du har blivit anlitad av ett skogsbolag för att undersöka antalet träd som fallit omkull/skadats efter en storm. Skogsbolaget har totalt 14712 hektar skog som är uppdelade i 413 områden som består av ett antal hektar skog vardera. Av dessa områden väljs 7 områden ut slumpmässigt för och antalet skadade och omkullfallna träd räknas inom varje område. Denna undersökning gav följande resultat:

	Antal hektar	Antal skador
1	32	455
2	29	395
3	34	462
4	37	588
5	34	532
6	35	546
7	34	522

   (a) Beräkna det totala antalet skadade på skogsbolagets hela bestånd, inklusve konfidensintervall (95 %). Använd den väntevärdesriktiga estimatorn, inte kvotestimatorn. **1p**.

   Svar: För att lösa denna uppgift behöver vi bara beräkna

$$\hat{t} = N \cdot \bar{y} = 413 \cdot 500 = 206500$$

   och

$$V(\hat{t}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s^2}{n} = 413^2 \cdot \left(1 - \frac{7}{413}\right) \frac{65.653^2}{7} = 1254.055^2$$

   vilket ger följande konfidensintervall:

$$\hat{t} \pm z_{\alpha/2} SE(\hat{t}) = 206500 \pm 1.96 \cdot 1254.055 \rightarrow [204042.051, 208957.949]$$

   (b) Beräkna kvoten antal skadade träd per hektar med tillhörande medelfel. **2.5p**.

   Svar: För att beräkna kvoten gör vi på följande sätt (se XX i Lohr) där antalet hektar är hjälpvariabeln x:

$$\hat{B} = \frac{\hat{t}}{\hat{t}_x} = \frac{206500}{13865} = 14.894$$

och för standardfelet använder vi XX i Lohr:

$$V(\hat{B}) = \left(1 - \frac{n}{N}\right) \frac{s_e^2}{\bar{x}_U \cdot n}$$

där

$$\bar{x}_U = \frac{t_x}{N} = \frac{14712}{413} = 35.622$$

För att beräkna s_e^2 behöver först $e_i = t_i - \hat{B}x_i$ beräknas vilket framgår i tabellen nedan. Med dessa uppgifter kan vi sedan beräkna $s_e^2 = 33.353^2$.

	x_i	t_i	e_i
1	32	455	-21.60
2	29	395	-36.91
3	34	462	-44.38
4	37	588	36.94
5	34	532	25.62
6	35	546	24.72
7	34	522	15.62

Sätter vi sedan in det i formeln ovan får vi:

$$V(\hat{B}) = \left(1 - \frac{n}{N}\right) \frac{s_e^2}{\bar{x}_U \cdot n} = \left(1 - \frac{7}{413}\right) \frac{33.353^2}{35.622 \cdot 7} = 0.363^2$$

(c) Beräkna igen det totala antalet skadade träd med kvotestimatorn med tillhörande medelfel (även om urvalet är så litet att det finns en risk för bias i skattningen). **1.5p**.

Svar: För att beräkna det totala antalet träd med kvotestimatorn använder vi XX i Lohr:

$$\hat{t}_r = \hat{B}t_x = 14.894 \cdot 14712 = 219114.894$$

med följande standardfel

$$V(\hat{t}_r) = V(\hat{B}t_x) = V(\hat{B}) \cdot t_x^2 = 0.363^2 \cdot 14712^2 = 5334.8^2$$

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.