Tentamen i Surveymetodik 732G26

Måns Magnusson

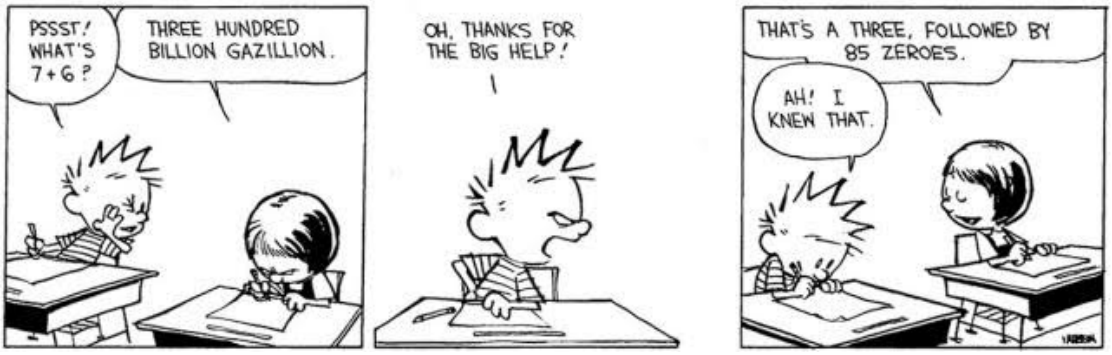
14 augusti 2013, kl. 8.00-12.00

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:**
  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.

Lycka till!



1. Inför med valet 2014 har en opinionsundersökning genomförts i bland den röstberättigade befolkningen i Motala kommun som består av 33657personer under sommaren 2013. Syftet är att med hjälp av dessa uppgifter undersöka den politiska opinionen och studera hur det kan tänkas gå i valet 2014. Från den röstberättigade befolkningen har 1000 personer valts ut med OSU och av dessa har 672 personer svarat. Av de 672 personer som svarat har 322 personer angett att de kommer rösta på Alliansen

   (a) Om syftet är att försöka förutsäga valet 2014. Vilka felkällor bedömmer du vara de största problemet för denna undersökning? **2p**.

   Svar: De felkällor som är uppenbara är:

   - Bortfallsfel (nästan 50 % bortfall)
   - Specifikationsfel (att med en undersökning 2013 försöka skatta hur många som kommer rösta i valet 2014 är osäkert, många kan ändra sig under tiden fram till valet)
   - Täckningsfel (den röstberättigade befolkningen 2013 kan ändras till 2014 - ex. kommer de som fyller 18 innan valet att rösta)

   (b) Baserat på uppgifterna ovan beräkna andelen som skulle rösta på det Alliansen 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:

   $$\hat{p} = 0.479$$

   $$\hat{V}(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1} = \left(1 - \frac{672}{33657}\right)\frac{0.479 \cdot 0.521}{671} \approx 0.019^2$$

   Dessa uppgifter används sedan för att beräkna konfidensintervallet:

   $$\hat{p} \pm z_{\alpha/2}SE(\hat{p}) = 0.479 \pm 1.96 \cdot 0.019 \rightarrow [0.442, 0.516]$$

   (c) Av resultatet ovan, skulle du säga att Alliansen skulle få egen majoritet? Varför, varför inte? **1p**.

   Svar: Eftersom konfidensintervallet täcker 0.5 är det inte möjligt att veta. De kan ha egen majoritet och på grund av slumpfelet får vi ett resultt som inte innebär egen majoritet.

2. Polismyndigheten är intresserade av att undersöka hur stor del av de anställda som i tjänsten varit utsatta för våld eller hot om våld som en del i arbetsmiljöarbetet. Polismyndigheten har därför valt att göra en undersökning bland poliserna i Sverige för att uppskatta omfattningen av problemet, tanken är att dra ett slumpmässigt urval av 400 anställda vid Polismyndigheten och samla in erfarenheter av våld och hot om våld. Vid Polismyndigheten finns två grupper anställda, dels civilanställda (8457) och dels poliser (19890). De civilanställda arbetar mestadels på kontor medan poliserna ofta kan arbeta ute

"i fält", inte sällan vid obekväma arbetstider. Därför har det beslutats att en webbenkät ska skickas till de civilanställda medan poliser ska intervjuas per telefon.
Från en tidigare undersökning gjort vid en av polismyndigheterna kom det fram att poliser utsatts för våld eller hot om våld vid i genomsnitt 4.045 tillfällen det senaste halvåret (med standardavvikelse 3.204), medan bland civilanställda var istället i genomsnitt 0.684 tillfällen (med standardavvikelse 0.555).

(a) Vilken allokering mellan poliser och civilanställda skulle du rekommendera för undersökningen? Varför? **1p**.

Svar: Eftersom det är olika kostnader för de olika strata och de olika strata har olika varians bör optimal allokering användas, där både kostnad, populationsstorlek och varians beaktas.

(b) Polismyndigheten väljer (oavsett din rekommendation) att använda Neymannallokering. Baserat på siffrorna från den tidigare undersökningen, beräkna vilken punktskattningen och varians den nya undersökningen kan tänkas få. **3p**.

Svar: Neymanallokering framgår i (3.14) i Lohr [2009, s. 89 f.]. För att kunna göra en allokering behöver anta att $s_1 = S_1$ och $s_2 = S_2$, vilkt ger följande allokering till strata 1 (poliser):

$$n_1 = \frac{N_1 S_1}{N_1 S_1 + N_2 S_2} n = \frac{19890 \cdot 3.204}{19890 \cdot 3.204 + 8457 \cdot 0.555} \cdot 400 = 372.56 \approx 373$$

Övriga urvalet hamnar sedan i strata 2 (civilanställda):

$$n_2 = n - n_1 = 27$$

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

Designvikten beräknas sedan (i enlighet med Lohr ???) på följande sätt:

$$w_1 = N_1/n_1 = (19890/373) = 53.324$$

$$w_2 = N_2/n_2 = (8457/27) = 313.222$$

3. I en kommun går 14726 barn på förskola och i kommunen finns totalt 143 förskolor där dessa barn går. Kommunen är intresserad av att uppskatta vilka kostnader föräldrarna har för barnen i förskolan (förskoleavgift, utflykter, matsäckar m.m.) per månad och drar ett slumpmässigt urval av 5 stycken förskolor och intervjuar där samtliga föräldrar. I tabellen nedan framgår resultatet:

(a) Baserat på dessa resultat beräkna den genomsnittliga förskolekostnaden per barn med tillhörande konfidensintervall (95 %). (Använd den estimator som är väntevärdesriktig/"unbiased") **2p**.

Svar: I denna uppgift rör det sig om ett enstegs klusterurval. För att beräkna denna uppgift med den vanliga "unbiased" estimatorn använder vi oss av (5.12) och (5.13) i

	M_i	$\bar{y}_i$	s_i
1	117	1061	105
2	95	770	84
3	105	637	105
4	104	1237	89
5	125	637	88

Table 1: F¡U+00F6¿r¡U+00E4¿ldrars kostnad f¡U+00F6¿r f¡U+00F6¿rskola

Lohr [2009, s. 179], så först behöver vi räkna ut totalen för respektive skola i urvalet genom $t_i = \bar{y}_i \cdot M_i$ vilket ger följande resultat:

	M_i	$\bar{y}_i$	s_i	t_i
1	117	1061	105	124137
2	95	770	84	73150
3	105	637	105	66885
4	104	1237	89	128648
5	125	637	88	79625

Med dessa uppgifter kan vi skatta totalen och medelvärdet på följande sätt:

$$\hat{t}_{unb} = \frac{N}{n}\sum t_i = \frac{143}{5} \cdot 472445 = 13511927$$

med tillhörande varians:

$$\hat{V}(\hat{t}_{unb}) = N^2\left(1 - \frac{n}{N}\right)\frac{s_t^2}{n} \approx 143^2\left(1 - \frac{5}{143}\right)\frac{29513.19^2}{5} = 1854123.704^2$$

Eftersom M_0 är känd (14726 barn i förskolan) är det enkelt att beräkna $\hat{\bar{y}}_{unb}$ och $Var\left(\hat{\bar{y}}_{unb}\right)$:

$$\hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{M_0} = \frac{13511927}{14726} = 917.556$$

$$Var\left(\hat{\bar{y}}_{unb}\right) = \frac{1}{M_0^2}Var\left(\hat{t}_{unb}\right) = 125.908^2$$

Med dessa resultat är det sedan möjligt att beräkna ett konfidensintervall för kostnaderna på följande sätt:

$$\hat{\bar{y}}_{unb} \pm z_{\alpha/2}SE(\hat{\bar{y}}_{unb}) = 917.556 \pm 1.96 \cdot 125.908 \rightarrow [670.776, 1164.336]$$

(b) Antag nu att undersökningen görs om och istället för att intervjua samtliga föräldrar intervjuas bara föräldrarna till 20, 20, 20, 20, 20 barn i varje klass. Beräkna på nytt en skattning av de genomsnittliga kostnaderna med tillhörande konfidensintervall baserat på denna nya urvalsdesign, men utgå från att vi fått samma resultat avseende $\bar{y}_i$ och s_i som i uppgift a) (Använd även här den estimator som är väntevärdesriktig/"unbiased"). Vilken urvalsdesign skulle du föredra och varför? **3p**.

Svar: Eftersom bara 20 föräldrar ska intervjuas behöver även variationen inom varje

kluster beaktas i uträkningarna. För detta används 5.24 i Lohr [2009, s. 185].

$$\hat{V}(\hat{t}_{unb}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_t^2}{n} + \frac{N}{n} \cdot \sum_{i=1}^{n} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{s_i^2}{m_i}$$

Den första delen av variansen har redan beräknats i uppgift a) ovan, det som återstår är den sista delen av variansen (inom kluster). I tabellen nedan har variansen inom varje kluster beräknats:

	M_i	$\bar{y}_i$	s_i	t_i	m_i	$M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{s_i^2}{m_i}$
1	117	1061	105	124137	20	59582.25
2	95	770	84	73150	20	29925.00
3	105	637	105	66885	20	46856.25
4	104	1237	89	128648	20	38875.20
5	125	637	88	79625	20	57750.00

Med dessa uppgifter är det enkelt att beräkna variansen för totalen:

$$\hat{V}(\hat{t}_{unb}) = 1854123.704^2 + \frac{143}{5} \cdot 232988.7 = 1854123.704^2 + 2581.371^2 = 1854125.501^2$$

vilket sedan kan beräknas för $\hat{\bar{y}}_{unb}$ på följande sätt

$$Var\left(\hat{\bar{y}}_{unb}\right) = \frac{1}{M_0^2} Var\left(\hat{t}_{unb}\right) = 125.908^2$$

Detta är i princip samma resultat som ovan då alla inom varje förskola undersöktes. Således blir konfidensintervallet detsamma och om det kostar extra mycket att intervjua fler inom varje skola så är det bättre att dra ett urval.

4. Du har fått i uppdrag att delta som statistiker i en studie gällande hushållens utgifter i en kommun. Sedan tidigare har ni (av kostnadsskäl) valt en urvalsstorlek på 1000 hushåll som dras slumpmässigt från samtliga 20300 hushåll i kommunen. Tidigare studier har visat att det finns en stor risk för att bortfallet kommer vara stort i denna studie som nu planeras.

   (a) Nämn tre saker du kan göra för att förebygga bortfall i denna studie **1p**

   Svar: Exempel på insatser är:

   - Erbjuda belöningar (gärna i förväg)
   - Kortare frågeformulär
   - Fyrfärgstryck
   - Rekommenderat brev
   - Förkontakter
   - Uppföljning

(b) Du har nu genomfört studien och med ett bortfall på hela 38 %. I urvalsramen finns hushållens inkomst att tillgå för bortfallsanalys och bortfallsuppföljning, vilket ger följande resultat:

	Bortfall	Svarande	Antal hush<U+00E5>ll
- 400 tkr	198	211	8762
400 + tkr	187	404	11538

Table 2: Resultat: Studie avseende hush¡U+00E5¿llens utgifter

Du genomför därför en enklare bortfallsanalys med ett χ^2-test:

```
Pearson's Chi-squared test with Yates' continuity correction

data:  answer and income
X-squared = 28, df = 1, p-value = 0.000000121
```

Vad drar du för slutsats om bortfallet? Vad gör du för antagande om bortfallet? **1p**

Svar: Bortfallet är inte slumpmässigt utan beror på hushållets inkomst. Vi kan eventuellt anta MAR, men eftersom det verkar som att bortfallet är korrelerat med inkomsten kan det även vara korrelerat med utgifterna varför det finns risk för NMAR.

(c) Dina uppdragsgivare menar (oavsett vad du anser) att materialet behöver kalibreras efter hushållens inkomst. Beräkna både design- och g-vikterna för de olika inkomstkategorierna. **2.5p**

Svar: Rent formellt är designvikten inversen av inklutionssannolikheten vilket ger

$$w = \frac{1}{\pi} = \frac{20300}{1000} = 20.3$$

men i detta fall kan vi beräkna designvikten baserat på det urval vi fått vilket ger

$$w = \frac{1}{\pi} = \frac{20300}{615} = 33.008$$

Att räkna upp vikterna på detta sätt innebär att vi (i detta steg antar MCAR), men självklart kan denna uppräkning även ligga "i" g-vikten.
Nästa steg är att beräkna g-vikten, vilket kan i detta fall beräknas på så sätt som anges i (4.12) i Lohr [2009, s. 132].

$$g_i = \frac{t_x}{\hat{t}_x}$$

vilket ger att för män

$$g_{-400tkr} = \frac{t_{-400tkr}}{\hat{t}_{-400tkr}} = \frac{8762}{\hat{p}_{-400tkr} \cdot N} = \frac{8762}{\frac{211}{615} \cdot 20300} = \frac{8762}{6964.715} = 1.258$$

och

$$g_{400tkr+} = \frac{t_{400tkr+}}{\hat{t}_{400tkr+}} = \frac{11538}{\hat{p}_{400tkr+} \cdot N} = \frac{11538}{\frac{404}{615} \cdot 20300} = \frac{11538}{13335.285} = 0.865$$

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.