Lösningar: Tentamen i Surveymetodik 732G26

Måns Magnusson

22 mars 2013, kl. 8.00-12.00

1. Det vi får givet i uppgiften är att n =600, $N = 91122$, $\bar{y}$=0.125, s =0.454 och $\hat{p}$ =0.093.

   (a) För att lösa denna uppgift använder vi oss av formel (2.15), (2.16) och (2.21) i Lohr [2009, s. 37, 42]. Detta ger:

   $$\hat{t} = N \cdot \bar{y} = 91122 \cdot 0.125 = 11390.25$$

   $$V(\hat{t}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s^2}{n} = 91122^2 \left(1 - \frac{600}{91122}\right) \frac{0.454^2}{600} = 1683.329^2$$

   $$\hat{t} \pm z_{\alpha/2} SE(\hat{t}) = 11390.25 \pm 1.96 \cdot 1683.329 \rightarrow [8090.986, 14689.514]$$

   (b) För att lösa denna uppgift använder vi oss av (2.19) i Lohr [2009, s. 38] för att beräkna variansen. Detta ger:

   $$\hat{p} = 0.093$$

   $$\hat{V}(\hat{p}) = \left(1 - \frac{n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1} = \left(1 - \frac{600}{91122}\right) \frac{0.093 \cdot 0.907}{599} \approx 0.012^2$$

   $$\hat{p} \pm z_{\alpha/2} SE(\hat{p}) = 0.093 \pm 1.96 \cdot 0.012 \rightarrow [0.07, 0.116]$$

   (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 av storleken $\bar{y} \pm$ 0.02. Detta innebär att $e = 0.02$ i detta fall. Vi behöver också anta att standardavvikelsen i populationen, S kan approximeras med standardavikelsen i vårt tidigare urval, s. Detta ger:

   $$n_0 = \left(\frac{z_{\alpha/2} S}{e}\right)^2 = \left(\frac{1.96 \cdot 0.454}{0.02}\right)^2 = 1979.465$$

   som sedan används för att beräkna det nya n:

   $$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{1979.465}{1 + \frac{1979.465}{91122}} = 1937.379 \approx 1938$$

2. Det vi får givet i denna uppgift är $N_1 = 36473$ vilket gör att vi kan beräkna N_2 till 54649. Vi har också att $\bar{y}_1 = 0.29$, $s_1 = 0.669$, $\bar{y}_2 = 0.031$ och s_2 =0.215 där strata 1 är gruppen med högre risk för arbetsskador och strata 2 är gruppen med lägre risk.

   (a) För att lösa denna uppgift använder vi formeln för stratifierade skattningar (3.1) och

(3.4) i Lohr [2009, s. 78 f.]. Detta ger att:

$$\hat{t}_{str} = \sum_{h=1}^{H} N_h \bar{y}_h = 36473 \cdot 0.29 + 54649 \cdot 0.031 = 12271.289$$

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

$$\left(1 - \frac{300}{36473}\right) 36473^2 \frac{0.669^2}{300} + \left(1 - \frac{300}{54649}\right) 54649^2 \frac{0.215^2}{300} = 1557.495^2$$

Konfidensintervallen beräknas som vanligt:

$$\hat{t}_{str} \pm z_{\alpha/2} SE(\hat{t}_{str}) = 12271.289 \pm 1.96 \cdot 1557.495 \rightarrow [9218.655, 15323.923]$$

Designeffekten kan nu enkelt beräknas som:

$$\frac{\hat{V}(\hat{t}_{str})}{\hat{V}(\hat{t})} = \frac{1557.495^2}{1683.329^2} = 0.856$$

(b) För att få en så bra skattning som möjligt använder vi oss av optimal allokering. Eftersom kostnaden är lika stor i respektive urval reduceras detta till Neymanallokering vilket framgår i (3.14) i Lohr [2009, s. 89 f.]. Vi behöver även här anta att $s_1 = S_1$ och $s_2 = S_2$. Detta ger att vi ska allokera följande antal till strata 1:

$$n_1 = \frac{N_1 S_1}{N_1 S_1 + N_2 S_2} n = \frac{36473 \cdot 0.669}{36473 \cdot 0.669 + 54649 \cdot 0.215} 600 \approx 405$$

och resten (d.v.s. 195) till strata 2.
Använder vi dessa stratumstorlekar för studien ovan får vi följande varians. Observera att $\hat{t}_{str}$ inte förändras.

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

$$\left(1 - \frac{405}{36473}\right) 36473^2 \frac{0.669^2}{405} + \left(1 - \frac{195}{54649}\right) 54649^2 \frac{0.215^2}{195} = 1469.416^2$$

Konfidensintervallen beräknas som vanligt:

$$\hat{t}_{str} \pm z_{\alpha/2} SE(\hat{t}_{str}) = 12271.289 \pm 1.96 \cdot 1469.416 \rightarrow [9391.286, 15151.292]$$

Designeffekten kan nu enkelt beräknas som:

$$\frac{\hat{V}(\hat{t}_{str})}{\hat{V}(\hat{t})} = \frac{1469.416^2}{1683.329^2} = 0.762$$

3. Det vi känner till sedan tidigare är att $N = 327$, M_0 =27012 och följande tabell:

	Anst<U+00E4>llda	Intervjuade	Medelv<U+00E4>rde	Standardavikelse
1	215	30	0.23	0.50
2	25	25	0.08	0.28
3	33	30	0.20	0.41
4	27	27	0.30	0.87
5	21	21	0.14	0.36
6	113	30	0.20	0.61

Table 1: Antal arbetsskador p¡U+00E5¿ de unders¡U+00F6¿kta arbetsplatserna

(a) Denna fråga var lite otydlig och kan beräknas på två sätt (båda ger rätt svar). Antingen med den "vanliga" unbiased estimatorn samt med en kvotestimator.
För att beräkna denna uppgift med den vanliga unbiased estimatorn använder vi oss av (5.12) och (5.13) i Lohr [2009, s. 179]. Detta ger:

$$\hat{t}_{unb} = \frac{N}{n}\sum t_i \approx \frac{327}{6}31.94 = 1740.73$$

där $t_i = \bar{y}_i \cdot M_i$ vilket ger följande varians

$$\hat{V}(\hat{t}_{unb}) = N^2\left(1 - \frac{n}{N}\right)\frac{s_t^2}{n} \approx 327^2\left(1 - \frac{6}{327}\right)\frac{5.567}{6} = 312.084^2$$

där s_t^2 frågår av Lohr [2009, s. 171].
För att beräkna genomsnittet behöver vi dividera både variansen och totalskattningen (se Lohr 2009, s. 179) med M_0 vilket ger följande skattningar:

$$\hat{\bar{y}}_{unb} = \frac{1}{M_0}\hat{t}_{unb} = 0.064$$

$$\hat{V}(\hat{\bar{y}}_{unb}) = \hat{V}(\hat{t}_{unb} \cdot \frac{1}{M_0}) = \frac{1}{M_0^2}\hat{V}(\hat{t}_{unb}) = 0.012^2$$

och konfidensintervallet

$$\hat{\bar{y}}_{unb} \pm z_{\alpha/2}SE(\hat{\bar{y}}_{unb}) = 0.064 \pm 1.96 \cdot 0.012 \rightarrow [0.042, 0.087]$$

Som nämdes ovan var frågan otydlig och även en kvotestimator kan användas för denna beräkning under antagandet att eftersom vi tittar på antalet intervjuade är M_0 okänd och måste skattas.
Först beräknar vi därför $\hat{\bar{y}}_r$ på följande sätt (5.15 i Lohr 2009):

$$\hat{\bar{y}}_r = \frac{\sum_{i\in\mathcal{S}} M_i\bar{y}_i}{\sum_{i\in\mathcal{S}} M_i} = \frac{31.94}{163} = 0.196$$

Sedan räknar vi ut variansen med (5.17) i Lohr [2009, s. 180]

$$\hat{V}(\hat{\bar{y}}_r) = \left(1 - \frac{n}{N}\right) \frac{1}{n\bar{M}^2} \frac{\sum_{i\in\mathcal{S}} M_i^2(\bar{y}_i - \hat{\bar{y}}_r)^2}{n-1} =$$

$$\left(1 - \frac{6}{327}\right) \frac{1}{6 \cdot 27.167^2} \frac{18.749}{5} = 0.029^2$$

Den stora skillnaden mellan tt använda ratioestimatorn och den vanliga "unbiased" estimatorn beror på skillnaden mellan att använda M_0 som är känd (vanliga estimatorn) eller att istället använda $\hat{M}_0$ (ratioestimatorn).

(b) I denna uppgift kan man även här använda två metoder, eftersom frågan inte var tydligt formulerad. Både kvotestimatorn och den vanliga unbiased estimatorn går bra att använda, även om kvotestimatorn är att föredra då skillnaden mellan de olika arbetsplatserna är så stora. Inledningsvis beräknar vi punktskattningen antingen genom som vanlig unbiased estimation eller med kvotestimation. Detta ger:

$$\hat{\bar{y}}_{unb} = \frac{1}{M_0}\hat{t}_{unb} = \frac{1}{M_0}\frac{N}{n}\sum_{i\in\mathcal{S}} M_i\bar{y}_i = \frac{1}{27012}\frac{327}{6}91.69 = 0.185$$

samt

$$\hat{\bar{y}}_r = \frac{\sum_{i\in\mathcal{S}} M_i\bar{y}_i}{\sum_{i\in\mathcal{S}} M_i} = \frac{91.69}{434} = 0.211$$

Variansen kan antingen lösas genom variansen för kvotestimatorn (5.28 i Lohr 2009, s. 185) eller variansen för den vanliga estimatorn (5.24 i Lohr [2009, s. 185]. Den del som är gemensam för båda metoderna är hur variansen beräknas inom varje kluster $\sum M_i^2 \left(1 - \frac{m}{M_i}\right) \frac{s_i^2}{m_i}$ så denna del räknas ut först. Låt oss kalla denna del för $InomVar$. Detta ger följande beräkning:

$$InomVar = \sum M_i^2 \left(1 - \frac{m}{M_i}\right) \frac{s_i^2}{m_i} = 331.46 + 0 + 0.55 + 0 + 0 + 116.33 = 448.344$$

Nu är vi klara för att beräkna vår varians. Först gör vi det med den vanliga unbiased variansestimatorn (5.24):

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

$$\hat{V}_{unb}(\hat{t}_{unb}) = 327^2 \left(1 - \frac{6}{327}\right) \frac{335.34}{6} + \frac{327}{6}\left(448.344\right) = 2427.147^2$$

Vi behöver sedan dividera denna varians med M_0 för att få $\hat{V}_{unb}(\hat{\bar{y}}_{unb})$:

$$\hat{V}_{unb}(\hat{\bar{y}}_{unb}) = \frac{1}{M_0^2}\hat{V}_{unb}(\hat{t}_{unb}) = 0.09^2$$

Vill vi istället använda oss av kvotestimatorn använder vi oss av (5.28) vilket ger:

$$\hat{V}_r(\hat{\bar{y}}_r) = \frac{1}{\bar{M}^2}\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{1}{nN\bar{M}^2}\cdot InomVar$$

där $\bar{M} = \frac{\sum M_i}{n} = 72.333$ och från (5.29):

$$s_r^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}} M_i^2(\bar{y}_i - \hat{\bar{y}}_r)^2 = \frac{1}{5}\cdot 36.729 = 7.346$$

Sammantaget ger detta resultatet

$$\hat{V}_r(\hat{\bar{y}}_r) = \frac{1}{72.333^2}\left(1-\frac{6}{327}\right)\frac{7.346}{6} + \frac{1}{6\cdot 327\cdot 72.333^2}\cdot 448.344 =$$

$$= 0.017^2$$

Vilket är betydligt mycket bättre än om den vanliga estimatorn används.

4. I denna uppgift har vi följande information att utgå ifrån:

	Bortfall	Svarande	Population	Antal (s)-v<U+00E4>ljare
M<U+00E4>n	160	232	50851	87
Kvinnor	158	255	51772	93

Table 2: Resultat: Opinionsunders¡U+00F6¿kning i Norrk¡U+00F6¿ping

(a) För att lösa denna uppgift använder vi oss av (2.19) i Lohr [2009, s. 38]för att beräkna variansen, observera att det är ett OSU och inte ett stratifierat urval. Detta ger (med N =102623):

$$\hat{p} = 0.37$$

$$\hat{V}(\hat{p}) = \left(1-\frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1} = \left(1-\frac{480}{102623}\right)\frac{0.37\cdot 0.63}{479} = 0.022^2$$

med konfidensintervallet

$$\hat{p} \pm z_{\alpha/2}SE(\hat{p}) = 0.37 \pm 1.96\cdot 0.022 \rightarrow [0.327, 0.412]$$

(b) Följande bortfallsanalys har gjorts:

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

data:  Resultat[, 1:2]
X-squared = 0.45, df = 1, p-value = 0.5026
```

Slutsatsen av denna (enkla) bortfallsanalys är att bortfallet inte beror på kön. Alltså kan vi anta MCAR. Dock är det tveksamt om bortfallet eventuelt beror på ålder (liksom andelen socialdemokrater, även om det inte testas här). Det gör att det eventuellt kan vara fråga om MAR, men vi har inte tillgång till variabeln ålder.

(c) Designvikten w_i påverkas (egentligen) inte av bortfallet. Men när vi fått ett urval med bortfall får vi utgå från att detta är vår urvalsstorlek och beräkna vikterna baserat på det urval vi fått.

$$w_R = \frac{N}{n_R} = \frac{102623}{487} = 210.725$$

g-vikten kan i detta fall beräknas som anges i (4.12) i Lohr [2009, s. 132].

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

vilket ger att för män

$$g_{män} = \frac{t_{män}}{\hat{t}_{män}} = \frac{50851}{\hat{p}_{män} \cdot N} = \frac{50851}{48888.164} = 1.04$$

och

$$g_{kvinnor} = \frac{t_{kvinnor}}{\hat{t}_{kvinnor}} = \frac{51772}{\hat{p}_{kvinnor} \cdot N} = \frac{51772}{53734.836} = 0.963$$

För att beräkna $\hat{t}$ används (4.12)

$$\hat{t}_{yr} = \sum w_i g_i y_i = w_i(g_{män} y_{män} + g_{män} y_{män}) = 210.725(1.04 \cdot 87 + 0.963 \cdot 93) =$$

$$= 38504.125$$

Vi jämför med resultatet i a) $\hat{p} \cdot N = 37930.472$ och konsterar att skilnaden är mycket liten i detta fall (vilket inte är så konstigt eftersom kön inte verkade förklara bortfallet).

References

S.L. Lohr. *Sampling: design and analysis.* Thomson, 2 edition, 2009.