Tentamen i Surveymetodik 732G26
Lösningsförslag

Måns Magnusson

21 mars 2014, kl. 8.00-12.00

Uppgift 1

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. Anltal svarande är $n_r = 772$. Detta ger:

$$\hat{t} = N\hat{p} = 1288778 \cdot 0.79 = 1018134.62$$

$$\begin{aligned}
\hat{V}(\hat{t}) &= \hat{V}(N \cdot \hat{p}) \\
&= N^2 \cdot \hat{V}(\hat{p}) \\
&= N^2 \cdot \left(1 - \frac{n_r}{N}\right) \frac{\hat{p}(1-\hat{p})}{n_r - 1} = \\
&= 1288778^2 \cdot \left(1 - \frac{772}{1288778}\right) \frac{0.79 \cdot 0.21}{771} \\
&\approx 18899.22537^2
\end{aligned}$$

Med detta är det sedan möjligt att beräkna konfidensintervallet

$$\begin{aligned}
\hat{t} \pm z_{\alpha/2} \cdot SE(\hat{t}) &= 1018134.62 \pm 1.645 \cdot 18899.22537 \\
&\rightarrow [987045.39426, 1049223.84574]
\end{aligned}$$

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 772 och den minsta urvalsstorleken som krävs för normalapproximation är enligt (2.23):

$$\begin{aligned}
n_{min} &= 28 + 25 \left(\frac{1}{s^3} \sum_{i \in \mathcal{S}} \frac{(y_i - \bar{y})^3}{n}\right)^2 \\
&= 28 + 25 \left(\frac{1}{2.6^3} 9.78\right)^2 \\
&= 28 + 25\,(0.55644)^2 \\
&= 28 + 7.74065 \\
&\approx 36
\end{aligned}$$

Det är således inga problem att använda normalapproximation i detta fall.

c) För att lösa denna uppgift används resultaten från Lohr [2009, kap. 2.4]. Observera att inklutionssannolikheten i designen beräknas på urvalstorleken (alla i populationen har samma sannolikhet att inkluderas i urvalet). Detta ger vid OSU att:

$$\pi_i = \frac{n}{N} = \frac{1250}{1288778} = 0.00097$$

Uppgift 2

a)

$$\begin{aligned}
\hat{V}(\hat{\bar{y}}) &= \left(1 - \frac{n}{N}\right)\frac{s^2}{n} \\
&= \left(1 - \frac{n}{N}\right)\frac{1}{n}\frac{\sum_i (y_i - \bar{y})^2}{n-1} \\
&= \left(1 - \frac{n}{N}\right)\frac{1}{n}\frac{\sum_i (y_i - \hat{p})^2}{n-1} \\
&= \left(1 - \frac{n}{N}\right)\frac{1}{n}\frac{\sum_i y_i^2 - 2\hat{p}\sum_i y_i + n\hat{p}^2)^2}{n-1} \\
&= \left(1 - \frac{n}{N}\right)\frac{1}{n}\frac{n\hat{p} - n2\hat{p}^2 + n\hat{p}^2}{n-1} \\
&= \left(1 - \frac{n}{N}\right)\frac{\hat{p} - \hat{p}^2}{n-1} \\
&= \left(1 - \frac{n}{N}\right)\frac{\hat{p}(1 - \hat{p})}{n-1} \\
&= \hat{V}(\hat{p})
\end{aligned}$$

d)

i) Sannolikheten att inkluderas i en undersökning är:

$$\pi_{JAG} = \frac{n}{N} = \frac{2000}{6000000} \approx 0.00033$$

ii) Sannolikheten att inte delta i någon av de 100 undersökningarna är:

$$(1 - 0.00033)^{100} \approx 0.96721$$

Uppgift 3

Se kurslitteraturen och föreläsningsanteckningar.

Uppgift 4

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

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

där resultatet för respektive strata framgår av tabellen nedan.

	N_h	$\bar{y}_h$	$\frac{N_h}{N} \cdot \bar{y}_h$
1	41361	3982.60	580.77
2	107289	3617.71	1368.47
3	134981	4299.55	2046.17

Detta ger att:

$$\begin{aligned} \hat{\bar{y}}_{str} &= 580.77 + 1368.47 + 2046.17 \\ &= 3995.41082 \end{aligned}$$

Sedan beräknas variansen med hjälp av:

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

Notera att vi bara har observerat (fått svar från) n_{hr} i respektive strata. Respektive del framgår av tabellen nedan:

	N_h	n_{rh}	s_h	$\left(\frac{N_h}{N}\right)^2$	$1 - \frac{n_{rh}}{N_h}$	$\frac{{s_h}^2}{n_{rh}}$	$\left(1 - \frac{n_{rh}}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{{s_h}^2}{n_{rh}}$
1	41361	93	166.86	0.02	1.00	299.38	6.35
2	107289	235	649.59	0.14	1.00	1795.60	256.37
3	134981	277	1414.89	0.23	1.00	7227.13	1633.47

Detta ger således att

$$\begin{aligned} \hat{V}(\hat{\bar{y}}_{str}) &= 6.35 + 256.37 + 1633.47 \\ &= 1896.18937 \end{aligned}$$

Konfidensintervallen kan sedan beräknas på följande sätt

$$\begin{aligned} \hat{\bar{y}}_{str} \pm z_{\alpha/2} \cdot \sqrt{\hat{V}(\hat{\bar{y}}_{str})} &= 3995.41082 \pm 2.576 \cdot 43.54526 \\ &\rightarrow [3883.23824, 4107.5834] \end{aligned}$$

b)

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	$\frac{N_h}{n_h}$
1	41361	146	283.29
2	107289	378	283.83
3	134981	476	283.57

c) I detta fall bör lika allokering användas. En lika allokering ger:

$$n_1 = ... = n_h = \frac{n}{H} = \frac{1000}{3} \approx 333$$

d) Då bör proportionell allokering användas.

References

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