Surveymetodik
Föreläsning 4

Måns Magnusson

Avd. Statistik, LiU

Översikt

Section 1

Stratifierat urval

Introduktion till stratifierat urval

- Ofta har vi tillgång till (kategoriska) hjälpvariabler innan vi drar vårt urval
  - Exempel: Kön, Ålder, Region o.s.v.
- Istället för att dra ett OSU kan vi stratifiera vår population i H **ömsesidigt uteslutande** "delpopulationer"
  - Exempel: Kön (2 klasser), Ålder (5 klasser) och Region (21 klasser)

$$= 2 \cdot 5 \cdot 21 = 210 \text{ strata}$$

- Från **varje strata** dras sedan (ofta) ett OSU
- Vi betraktar de olika strata som **olika populationer**.
- **Obs!** strata $\neq$ redovisningsgrupp

Introduktion till stratifierat urval II

- Varför vill vi stratifiera? Fler anledningar finns, nämligen...
  - Försäkra oss om att **inte få ett dåligt urval** (av en slump)
  - Vi vill ha en given precision för en eller flera **redovisningsgrupper**
  - Ibland har olika strata **olika kostnad**
  - Med stratifierade urval kan vi **öka den totala precisionen**
  - Olika grupper kan ha olika **bortfall**
- Det är inte ovanligt att flera av dessa faktorer spelar in på en och samma gång.
- Vanliga begrepp vid stratifierade urval
  - Take-all strata
  - Cut-off sampling

Subsection 1

Notation

Notation

$$N = \text{populationsstorlek}$$

$$N_h = \text{populationsstorlek i strata } h$$

$$\mathcal{U}_h = \text{populationsmängden i strata } h$$

$$N = N_1 + N_2 + ... + N_h = \sum_{i=1}^{H} N_i$$

$$n = \text{urvalsstorlek}$$

$$n_h = \text{populationsstorlek i strata } h$$

$$\mathcal{S}_h = \text{urvalsmängden i strata } h$$

$$n = n_1 + n_2 + ... + n_h = \sum_{i=1}^{H} n_i$$

$$y_{hj} = \text{variabelvärde på enhet } j \text{ i stratum } h$$

Populationsparametrar

$$y_{hj} = \text{variabelvärde på enhet } j \text{ i stratum } h$$

$$t_{hy} = \sum_{j=1}^{N_h} y_{hj} = \text{total i stratum } h$$

$$t_y = t_{1y} + t_{2y} + ... + t_{Hy} = \sum_{h=1}^{H} t_{hy}$$

$$\bar{y}_{h\mathcal{U}} = \frac{1}{N_h} \sum_{j=1}^{N_h} y_{hj} = \text{medelvärde i strata } h$$

$$\bar{y}_{\mathcal{U}} = \frac{1}{N} \sum_{h=1}^{H} \sum_{j=1}^{N_h} y_{hj} = \frac{1}{N} \sum_{h=1}^{H} N_h \bar{y}_{h\mathcal{U}} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{h\mathcal{U}}$$

Vi ser också att

$$t = N\bar{y}_{\mathcal{U}} = \sum_{h=1}^{H} N_h \bar{y}_{h\mathcal{U}}$$

Subsection 2

Estimation

Estimation

- I varje enskilt strata - som vid vanligt OSU
- Sedan kombineras dessa till en “helhets”-skattning

- Vi kan exempelvis använda oss av den “vanliga” estimatorn $\bar{y}_{\mathcal{S}}$ i varje strata h

$$\bar{y}_h = \frac{1}{n_h} \sum_{j \in \mathcal{S}_h} y_{hj}$$

- Sedan lägger vi ihop (summerar) de enskilda skattningarna för respektive strata till en skattning för hela populationen

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

- Om $\bar{y}_h$ är väntevärdesriktig är också $\bar{y}_{str}$ och $\hat{t}_{str}$ väntevärdesriktiga $(*)$ (för härledning se Lohr, 2009, s. 79).

Estimation II

- $\hat{Var}(\bar{y}_{str})$ är i summan av variansen över de olika strata $(*)$

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

- Detta innebär:
  - Minst två observationer i varje strata
  - Vi ökar precisionen vid stratifiering om det är **stor variation mellan strata och liten variation inom strata**.
  - Det går att använda andra estimatorer (kvot- regressionsestimation)

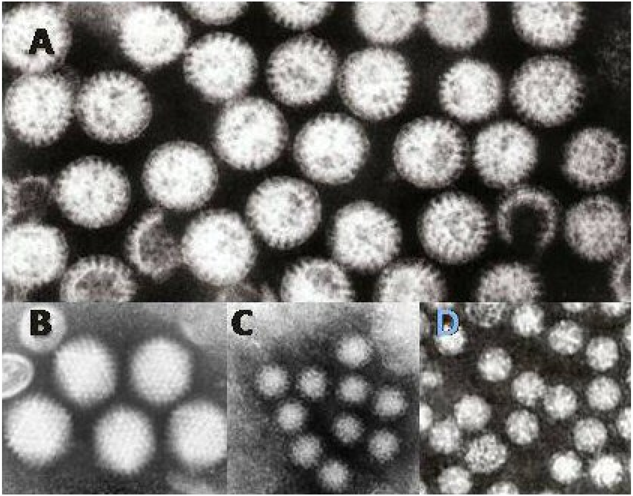
Estimation III

- För att beräkna ett konfidensintervall används (som vanligt) följande formel

$$\bar{y}_{str} \pm z_{\alpha/2}\sqrt{\hat{Var}(\bar{y}_{str})} \text{ och } \hat{t}_{str} \pm z_{\alpha/2}\sqrt{\hat{Var}(\hat{t}_{str})}$$

- I ex. SAS eller R används ibland $t_{\alpha/2,n-H}$ istället för $z_{\alpha/2}$
- För proportione p_{str} se Lohr (2009, s. 80 f.)

Exempel: Gastroenterit

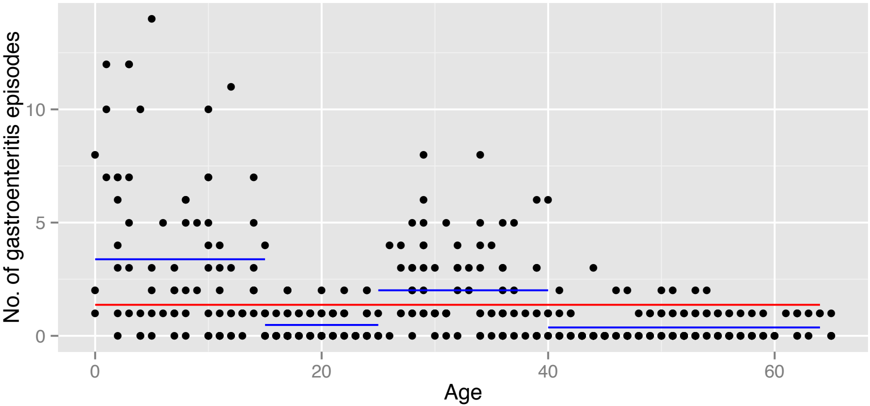


Figur : Gastroenteritvirus: A = rotavirus, B = adenovirus, C = norovirus och D = astrovirus.
Källa: Wikipedia

Exempel: Gastroenterit

- Vi är intresserade av att undersöka antalet fall av gastroenterit (magsjuka) i Sverige under 2011 i befolkningen 0-64 år.
- Vi är intresserade av det totala fallen av magsjuka, men också antalet fall bland yngre barn.
- Vi vet också sedan tidigare att småbarnsföräldrar är mer magsjuka än övriga åldersgrupper.
  - Vi skapar därför fyra strata (0-15 år, 16-25 år, 26-40 år och 41-65 år)
  - Vårt urval är på $n = 400$
- Urvalet fördelas proportionellt

Exempel: Gastroenterit II



Exempel: Gastroenterit III

Strata	N_h	n_h	$\bar{y}_h$	s_h
0-15 år	1687283	81	3.38	3.5
16-25 år	1851959	89	0.48	0.64
26-40 år	1812691	88	2.01	1.91
41-64 år	2935231	142	0.37	0.61
Samtliga	8287164	400	1.37	2.22

- Vad är det totala antalet fall av gastroenterit i befolkningen med tillhörande konfidensintervall?
- Hur stort urval skulle vi behöva för att få samma precision med vanligt OSU?

Exempel: Gastroenterit, lösning

- Först beräknar vi $\bar{y}_{str}$ Först beräknar vi $\bar{y}_{str}$

$$\bar{y}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = 1.37$$

- Sedan beräknar vi medelfelet för $\bar{y}_{str}$

$$SE(\bar{y}_{str}) = \sqrt{\frac{1}{N^2} \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}} = 0.094$$

- Vi kan jämföra detta med om vi använt vanligt OSU

$$\bar{y}_{\mathcal{S}} = 1.3675$$

$$SE(\bar{y}_{\mathcal{S}}) = \sqrt{\frac{s^2}{n} \left(1 - \frac{n}{N}\right)} = 0.111$$

- För att få samma precision med vanligt OSU krävs ett urval på $n \approx 557$

Subsection 3

Design effect

Design effect

- Som vi såg i exemplet kan vi tjäna på att stratifiera vårt urval
- Hur mycket vi tjänar brukar uttryckas som **design effect**

$$deff_\theta = \frac{Var(\hat{\theta} \mid \textbf{vår urvalsplan})}{Var(\hat{\theta} \mid \textbf{OSU})}$$

där $\hat{\theta}$ är vår estimator.

- Tumregel:
  - Om $deff_\theta < 1$ är studien **mer** precis än OSU (ex. stratifierat urval)
  - Om $deff_\theta > 1$ är studien **mindre** precis än OSU (ex. klusterurval)
- $deff_\theta$ går ofta att få från tidigare studier (vid ny design av studie)
- Exempel: Vad var designeffekten i Gastroenteritstudien i exemplet ovan?

Exempel: Design effect, lösning

- Vår estimator är skattningen av medelvärdet:

$$\theta = \hat{\bar{y}}_{\mathcal{U}}$$

vilket ger följande design effect

$$\begin{aligned} deff_{\theta} = \frac{Var(\hat{\bar{y}}_{\mathcal{U}} \mid \textbf{vår urvalsplan})}{Var(\hat{\bar{y}}_{\mathcal{U}} \mid \textbf{OSU})} &= \frac{Var(\bar{y}_{str})}{Var(\bar{y}_{\mathcal{S}})} \\ &= \frac{0.008836}{0.012321} \\ &= 0.717 \end{aligned}$$

Subsection 4

Allokering av urval till strata

- Det finns många sätt att allokera sitt urval till de olika strata
  De vanligaste allokeringsmetoderna är:
  - Lika allokering
  - Proportionell allokering
  - Optimal- och Neymanallokering
  - (Powerallokering)

Allokering av urval - Exempel

- Vi använder Gastroenteritexemplet ovan.

Strata	N_h	$\bar{y}_h$	s_h
0-15 år	1687283	3.38	3.5
16-25 år	1851959	0.48	0.64
26-40 år	1812691	2.01	1.91
41-64 år	2935231	0.37	0.61
Samtliga	8287164	1.37	2.22

- Vi antar (initialt) att kostnaden för datainsamlingen (c_h) skulle vara lika stor i alla strata.
- Hur skulle detta urval allokeras med de olika sätten att allokera urval?

Lika allokering

- Urvalet allokeras till jämt fördelat över alla strata

$$n_1 = n_2 = ... = n_H = \frac{n}{H}$$

- För- och nackdelar:
  + Enkel design
  - Ineffektiv

- Vad blir detta i exemplet ovan?

Proportionell allokering

- Urvalet allokeras enligt proportionellt mot de olika stratastorlekarna

$$\frac{n_h}{n} = \frac{N_h}{N} \text{ för } h = 1, 2, ..., H$$

- För- och nackdelar:
  + Enkel design
  + "Självvägande" urval (minimal variation i w_i)

- Vad blir detta i exemplet ovan?

Optimal allokering

- Urvalet fördelas så att $Var(\bar{y}_{str})$ minimeras (givet olika kostnader och varians i olika strata)

$$n_h = n \cdot \frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}}}$$

där

$$c_h = \text{kostnaden per observation i strata } h$$

- Om kostnaden är lika i respektive strata kallas allokeringen för **Neymanallokering**

$$n_h = n \cdot \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}$$

- Om även variansen är lika i respektive strata reduceras Neymanallokering till **proportionell** allokering

Optimal allokering II

- För- och nackdelar:
  + Optimal design (mindre n)
  + Kan ta kostnader i beaktande
  - Optimerar efter endast en variabel
  - Tar inte hänsyn till intresse av redovisningsgrupper
- Exempel på användning av kostnader i olika strata:
  - Ta hänsyn till olika bortfall
  - Ta hänsyn till resekostnader vid klusterurval
- Vad blir detta i exemplet ovan (om vi antar lika kostnader)?

Powerallokering

- Ofta intresse av **både** hela populationen **och** delpopulationer (redovisningsgrupper)
- Om vi vill göra en avvägning mellan en optimal sammanlagd skattning och skattningar i reovisningrupperna $\rightarrow$ **powerallokering**
- Ibland enklare att bestämma precisionen för respektive redovisningsgrupp $\rightarrow$ den totala precisionen blir då mindre mindre

Powerallokering II

- För att göra en powerallokering används följande formel (se Bankier (1988)):

$$n_h = n \cdot \frac{S_h X_h^q / \bar{y}_{h\mathcal{U}}}{\sum_{h=1}^{H} S_h X_h^q / \bar{y}_{h\mathcal{U}}}$$

  där X_h är vilken "betydelse" vi tillskriver strata h och q är en konstant på intervallet $0 \leq q \leq 1$.

- Sätter vi $X_h = t_{hy}$ så resulterar $q = 1$ i Neymanallokering och $q = 0$ resulterar i approximativt samma precision i alla strata.
- Ett vanligt värde som kan användas är $q = \frac{1}{2}$ eller $q = \frac{1}{3}$ (se Lehtonen et al. (2004, s. 65))

Design effect för olika urvalsallokering

Allokering	n_1	n_2	n_3	n_4	$Var(\hat{\bar{y}}_{\mathcal{U}})$	$deff(\hat{\bar{y}}_{\mathcal{U}})$
OSU	-	-	-	-	0.111	1
Lika	100	100	100	100	0.087	0.614
Proportionell	81	89	88	142	0.094	0.717
Neyman	191	39	112	58	0.075	0.457

- För N_h, s_h^2 och $\bar{y}_h$ i respektive strata se sida 23.

Subsection 5

Urvalsvikter vid stratifiering

Urvalsvikter vid OSU

- Designvikten **vid OSU** är

$$w_i = \frac{1}{\pi_i} = \frac{N}{n}$$

- Vikter vid estimation $(*)$

$$\hat{t}_{\mathcal{U}} = \hat{t}_{HT} = \sum_{i \in \mathcal{S}} w_i y_i = \sum_{i \in \mathcal{S}} \frac{y_i}{\pi_i}$$

$$\hat{\bar{y}}_{\mathcal{U}} = \frac{\hat{t}_{\mathcal{U}}}{N} = \frac{\sum_{i \in \mathcal{S}} w_i y_i}{\sum_{i \in \mathcal{S}} w_i}$$

då

$$\sum_{i=1}^{N} \pi_i = n \text{ och } \sum_{i=1}^{n} w_i = N$$

- Denna estimator kallas **Horwitz-Thompson** estimatorn.
- Detta går enkelt att vidareutveckla för stratifierade urval

Urvalsvikter vid stratifiering

- För ett stratifierat urval får vi samma formel för totalskattningen $(*)$:

$$\hat{t}_{\mathcal{U},str} = \hat{t}_{HT} = \sum_{j \in \mathcal{S}} w_j y_j$$

- För att beräkna $\hat{\bar{y}}_{\mathcal{U},str}$ används de "stratifierade" vikterna.

$$\hat{\bar{y}}_{\mathcal{U}} = \frac{\hat{t}_{\mathcal{U}}}{N} = \frac{\sum_{j \in \mathcal{S}} w_j y_{hj}}{\sum_{j \in \mathcal{S}} w_j}$$

  Det går att visa att denna estimator är samma som estimatorn på sida 10 $(*)$.

- Vid proportionell allokering har vi ett specialfall

$$w_{hj} = \frac{N_h}{n_h} = \frac{N}{n}$$

  vilket innebär att vi får samma vikter i alla strata.
- Då w_j inte beror på strata kallas urvalet "självvägt"

Urvalsvikter, Exempel

- Utgå från Gatsroenteritexemplet på sida 23 och 30.
- Beräkna med Horwitz-Thompson estimatorn en totalskattning vid proportionellt urval och vid Neymanurval.

Referenser

Bankier, M., 1988. Power allocations: determining sample sizes for subnational areas. The American Statistician 42 (3), 174–177.

Lehtonen, R., Pahkinen, E., Wiley, J., 2004. Practical methods for design and analysis of complex surveys. J. Wiley.

Lohr, S., 2009. Sampling: design and analysis, 2nd Edition. Thomson.

Wikipedia, ???? gastroenteritis viruses".