Surveymetodik
Föreläsning 11

Måns Magnusson

Avd. Statistik, LiU

Översikt

1 Klusterurval
- Enstegs klusterurval
- Systematiskt urval

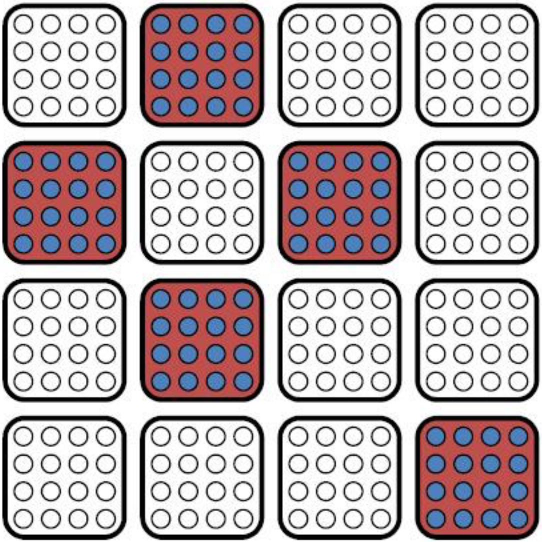
Section 1

Klusterurval

Introduktion till klusterurval

- Ibland finns naturliga grupper/kluster i populationen
  - **Exempel:** Skolor, Bostadsområden, Arbetsplatser, Hushåll
- I ett klusterurval dras ett urval av...
  - kluster (primära urvalsenheter, **psu**) och
  - ett antal observationsenheter i varje kluster (sekundära enheter, **ssu**)

Introduktion till klusterurval II



(a) Cluster sampling with $m = 5$

Jämförelse stratifierat urval och klusterurval

- Stratifierat urval
  - **Mål:** Stor skillnad mellan strata, liten inom strata
  - Samtliga strata "representerar" **sin del** av målpopulationen
  - Vi tar ett OSU i **varje** strata
  - Ju större skillnaden mellan strata ju **mindre** design effect
  - **Maximal** deff $= 1$ (lite grovt)
- Klusterurval
  - **Mål:** Liten skillnad mellan kluster, stor skillnad inom kluster
  - Samtliga kluster "representerar" **övriga** kluster/målpopulationen
  - Vi tar ett OSU i **ett urval** av kluster
  - Ju större skillnad mellan kluster ju **större** design effect
  - **Minimal** deff $= 1$ (lite grovt)

- Varför vill vi använda klusterurval?
  - Det är svårt eller omöjligt att konstruera en urvalsram
  - Det förekommer naturliga kluster vilket minskar insamlingskostnaden
  - Andra metodologiska fördelar (ex. lägre bortfall)
- Varför vi vill undvika klusterurval?
  - Nästan alltid sämre design effect än vid OSU
  - Mer komplicerade beräkningar

Exempel: Attityder till skolan

Attityder till skolan (Skolverket, 2010)
Skolverket

- **Syfte:** Hur situationen ser ut för elever och lärare i skolan.
- **Målpopulation:** Yngre elever (årskurs 4-6) och äldre elever (årskurs 7-9 och gymnaisiet).
- **Urval:**
  Yngre elever: 2 645 respondenter (155 skolor), Tvåstegs klusterurval
  Äldre elever: 2 600 respondenter, Stratifierat OSU
- **Bortfall:**
  Yngre elever: 5 % av skolorna, 8 % av eleverna
  Äldre elever: 27 %
- **Datainsamlingsmetod:** Pappersenkäter i klassrummet (yngre elever) och telefonintervjuer (äldre elever)
- **Periodicitet**: Varje 3:e år.

Notation och begrepp I

- För att förtydliga exemplifierar vi med ett klusterurval av skolor där
  - Skolor är den primära urvalsenheten (**psu**)
  - Elever är den sekundära urvalsenheten (**ssu**)
  - Vi är intresserad av antalet sjukdagar (y)

- Notation på **psu**-nivå (skolnivå) - populationsparametrar

$$N = \text{antal psus (skolor) i populationen}$$

$$M_i = \text{antal ssus (elever) i psu (skola) } i$$

$$M_0 = \text{antal ssus (elever) i populationen}$$

$$y_{ij} = \text{observation i skola } i \text{avseende elev } j$$

Notation och begrepp II

$$t_i = \sum_{j=1}^{M_i} y_{ij} = \text{total i psu (skola) } i$$

$$t = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} \text{ populationstotal}$$

$$S_t^2 = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} \text{ populationsvarians av psu totaler}$$

Notation och begrepp III

$$\bar{y}_{\mathcal{U}} = \frac{1}{M_0} \sum_{i,j} y_{ij} = \text{populationsmedelvärde avseende } y \text{i populationen}$$

$$\bar{y}_{i\mathcal{U}} = \frac{1}{M_i} \sum_{i,j} y_{ij} = \frac{t_i}{M_i} = \text{populationsmedelvärde avseende } y \text{i psu (skola) } i$$

$$S^2 = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_{\mathcal{U}})^2}{M_0 - 1} = \text{populationsvarians avseende } y$$

$$S_i^2 = \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_{i\mathcal{U}})^2}{M_i - 1} = \text{populationsvarians avseende } y \text{i psu (skola) } i$$

Notation och begrepp IV

- Två typer av populationsmedel/total:

$$t, t_i \text{ och } \bar{y}_{\mathcal{U}}, \bar{y}_{i\mathcal{U}}$$

- Tre typer av populationsvarians

$$S^2, S_t^2, S_i^2$$

Exempel

- Ett mejeri har ett antal lastbilar som kör ut mjölkprodukter för leverans.
- Mjölkpaketen är packade i pallar som kan ta 200 mjölkpaket vardera.
- Totalt levererar mejeriet 1200 pallar mjölk per månad till olika livsmedelsbutiker.
- Mejeriet är intresserade av hur många mjölkpaket som skadas vid transporten.

- Vad är $N, M_i, M_0, t_i, t, y_{ij}, S_t^2$ i detta fall?

Olika typer av klusterurval

- Enstegs klusterurval
  - Vi observerar alla ssu (elever) i de psu (skolor) som dras i urvalet
  - $M_i = m_i, t_i = \hat{t}_i, S_i^2 = s_i^2$
- Flerstegs (exempelvis två) klusterurval
  - Vi ett urval av ssu (elever) i de psu (skolor) som dras i urvalet
  - $M_i > m_i$
- Som vid stratifiering så kan kluster ses som "egna" subpopulationer
- Kombineras till en "gemensam" skattning

Notation och begrepp: Urval

- Notation för vårt urval:

$$n = \text{antal psus (skolor) i urvalet}$$

$$m_i = \text{antal ssus (elever) i psu (skola) } i \text{ i vårt urval}$$

$$m_0 = \text{antal ssus (elever) i vårt urval}$$

$$\hat{t}_i = \text{skattad total i psu (skola) } i$$

$$\bar{t} = \sum_{i=1}^{n} t_i = \text{genomsnittlig klustertotal i urvalet}$$

Notation och begrepp: Urval II

$\hat{t}_{unb}$ = 'unbiased' skattning av populationstotalen t

s_t^2 = varians i urvalet mellan psu (skolor)

s_i^2 = varians i urvalet inom psu (skola) i

w_{ij} = urvalsvikt i psu (skola) $j$avseende ssu (elev) i

Exempel II

- Vi ska göra en undersökning avseende inkomst i hus (parhus, radhus och villa) och väljer att dra tre slumpmässiga hus.
- Vi får följande resultat:
  - Hus 1: Göran med dotterna Åsa (inkomst 21 200) samt Lisa, Kajsa och sönerna Bill och Bull (inkomst 21 300 och 29800)
  - Hus 2: Tommy, Gunnar och Gunvor (inkomst 8100, 21200 och 29800) samt Yasmine och Cletus (inkomst 43 200 och 37 000)
  - Hus 3: Göran och Inga med dottern Tuva (inkomst 15 100 och 19 200)
- Vad innebär $n, m_i, m_0, y_{ij}, t_i, s_t^2, s_i^2, s^2$ i detta fall?

Subsection 1

Enstegs klusterurval

Enstegs klusterurval

- Det finns (ännu) inget urval inom klustret, alla urvalsenheter observeras.

- Precis som vanligt OSU - men istället är observationerna klustertotaler.
  - Ex. $t_1, t_2, ..., t_n =$ antal sjukdagar i skola $1, 2, ..., n$

- Som vid vanligt OSU kan vi skatta hela populationstotalen

$$\hat{t}_{\mathcal{U}} = \hat{t}_{unb} = N \cdot \bar{t} = \frac{N}{n} \sum_{i \in \mathcal{S}} t_i$$

(det var i princip detta vi gjorde med brott per område tidigare)

- Denna skattning är (precis som vid OSU) väntevärdesriktig

Enstegs klusterurval II

- Vi beräknar variansen (som vid ett vanligt OSU) men med t_i (istället för y_i)

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

där

$$s_t^2 = \frac{\sum_{i\in\mathcal{S}}(t_i - \bar{t})}{n-1} \text{ och } \bar{t} = \frac{1}{n}\sum_{i\in\mathcal{S}} t_i$$

- Konfidensintervall för $\hat{t}_{unb}$ beräknas (som vanligt):

$$\hat{t}_{unb} \pm z_{\alpha/2}\sqrt{Var(\hat{t}_{unb})}$$

Precis som vanligt OSU, men med t_i istället för y_i

Enstegs klusterurval III

- För att skatta $\bar{y}_{\mathcal{U}}$ använder vi oss av totalskattningen $\hat{t}_{unb}$ och det totala antalet element M_0

$$\hat{\bar{y}}_{\mathcal{U}} = \frac{\hat{t}_{unb}}{M_0}$$

- På samma sätt beräknar vi variansen $(*)$

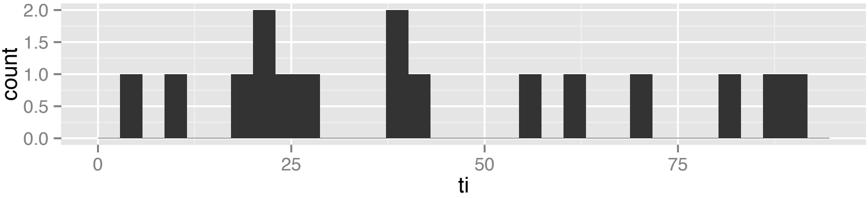
$$Var(\hat{\bar{y}}_{unb}) = \left(\frac{N}{M_0}\right)^2 \cdot \frac{s_t^2}{n}\left(1 - \frac{n}{N}\right)$$

- Konfidensintervallet beräknar vi (som vanligt):

$$\hat{\bar{y}}_{unb} \pm z_{\alpha/2}\sqrt{Var(\hat{\bar{y}}_{unb})}$$

Exempel: Barns sjukdagar

- Vi vill uppskatta det genomsnittliga antalet sjukdagar per barn och vecka. Vi drar ett klusterurval av 16 förskolor i Sverige år 2011.
- Det finns totalt $N = 10033$ förskolor och $M_0 = 472161$ barn.



- $t_i = (86, 41, 38, 82, 28, 11, 5, 22, 91, 56, 25, 63, 22, 38, 69, 20)$
- Vi vill räkna ut genomsnittligt antal sjukdagar per barn och vecka med tillhörande konfidensintervall.

- Denna skattning har två problem
  - t_i är troligtvis korrelerat med M_i
  - Vi kanske inte känner M_0
- Vi är intresserade av storheten

$$\hat{\bar{y}}_{\mathcal{U}} = \frac{\hat{t}_{unb}}{M_0}$$

som är en kvotskattning (av $\hat{B}$)

- Precis som vid vanlig kvotskattning:
  - antingen känner vi till M_0 eller inte.

- Skattning av kvoten är

$$\hat{B} = \frac{\bar{y}_{\mathcal{S}}}{\bar{x}_{\mathcal{S}}} = \frac{\sum_{i \in \mathcal{S}} y_i}{\sum_{i \in \mathcal{S}} x_i}$$

- Variansen för $\hat{B}$ skattas (se Lohr (2009, s. 125 f.) för detaljer)

$$\hat{Var}(\hat{B}) = \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n\bar{x}_{\mathcal{S}}^2}$$

- Om vi **känner till** $\bar{x}_{\mathcal{U}}$ och t_x så kan vi använda detta för att få mindre varians när vi skattar $\hat{\bar{y}}_r$ och $\hat{t}_{r,y}$ $(*)$

$$\hat{Var}(\hat{\bar{y}}_r) = \left(1 - \frac{n}{N}\right) \left(\frac{\bar{x}_{\mathcal{U}}}{\bar{x}_{\mathcal{S}}}\right)^2 \frac{s_e^2}{n}$$

Enstegs klusterurval V

- Byter vi ut y_i mot t_i och x_i mot M_i får vi att

$$\hat{B} = \hat{\bar{y}}_r = \frac{\hat{t}_{unb}}{\hat{M}_0} = \frac{\sum_{i \in \mathcal{S}} t_i}{\sum_{i \in \mathcal{S}} M_i}$$

och

$$\hat{Var}(\hat{\bar{y}}_r) = \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n \bar{M}_{\mathcal{S}}^2}$$

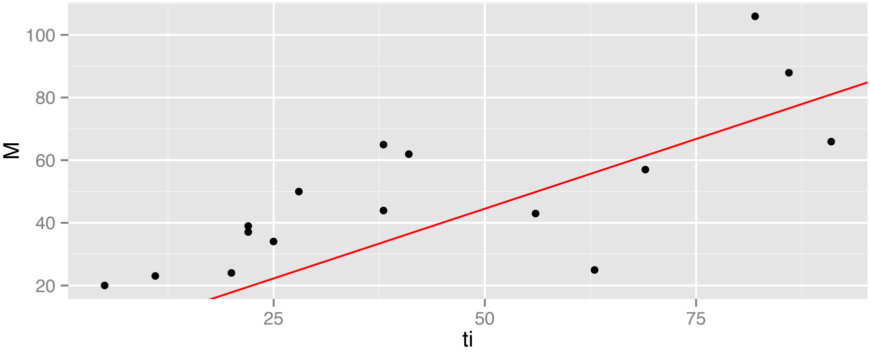
där

$$s_e^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} e_i^2 \text{ och } e_i = t_i - \hat{\bar{y}}_{\mathcal{U}} M_i$$

- Känner vi till M_0 så kan vi använda detta för att får mindre varians när vi skattar $\hat{t}_r$ (som vid vanlig kvotestimation)

$$\hat{Var}(\hat{t}_r) = \hat{Var}(\hat{\bar{y}}_r M_0) = M_0^2 \hat{Var}(\hat{\bar{y}}_r) = \left(1 - \frac{n}{N}\right) \left(\frac{M_0}{\bar{M}}\right)^2 \frac{s_e^2}{n}$$

Exempel: Barns sjukdagar (kvotskattning)



- Korrelationen är 0.75 och det är rimligt att om $M = 0$ så är $t_i = 0$.
- I urvalet har förskolorna följande storlek: $M_i = (88, 62, 65, 106, 50, 23, 20, 37, 66, 43, 34, 25, 39, 44, 57, 24)$
- Beräkna en skattning av det totala antalet sjukdagar med kvotestimatorn. $(*)$

Subsection 2

Systematiskt urval

Systematiskt urval

- Ett specialfall av (enstegs) klusterurval är systematiskt urval
- Används ofta...
  - det inte finns någon urvalsram, men hela populationen "passerar"
  - i äldre undersökningar (när det var enklare än att dra OSU)
  - när det är otympligt att dra ett OSU av praktiska skäl
    - "time-location" sampling
- Om det inte finns "periodicitet" i ramen så är

$$\text{Systematiskt urval} \approx \text{OSU}$$

Att dra ett systematiskt urval

- Vi känner (eller uppskattar) populationsstorleken N och vill ha en urvalsstorlek av storlek n
- Vi beräknar heltalet k på följande sätt

$$\left\lfloor \frac{N}{n} \right\rfloor = k$$

- Populationen $\mathcal{U}$ är nu indelad i k kluster
- Sedan väljer vi slumpmässigt en siffra, l, mellan 1 och k och väljer vart l:te urvalsenhet
  - I praktiken har vi dragit **ett** slumpmässigt kluster
- Vi får då följande urval

$$\mathcal{S} = \{l, l+k, l+2k, ..., l+(n-1)k\}$$

- Då k är en avrundning kan man behöva lägga till en urvalsenhet så att urvalsstorleken blir $n+1$
  - Så att alla element har en chans att bli dragna

Systematiskt urval III

- Då antalet kluster i urvalet bara är $n = 1$ så innebär det att det inte går att beräkna variansen mellan kluster
- En enkel lösning är att istället ta 2 (eller flera) kluster och beräkna k på följande sätt
$$2 \cdot \left\lfloor \frac{N}{n} \right\rfloor = k$$
och sedan slumpa två värden (kluster) l_1 och l_2 utan återläggning
- Då blir urvalet istället
$$\mathcal{S} = \{l_1, l_2, l_1 + k, l_2 + k, l_1 + 2k, l_2 + 2k, ...\}$$
och vi kan beräkna variansen korrekt (och inkludera/hantera eventuell periodicitet)

Exempel: Systematiskt urval

- Ett bolag som som säljer godis över internet är intresserade av kundnöjdheten hos sina kunder.
- De uppskattar att de har cirka $N = 12000$ kunder per månad på sin webbplats och är intresserad av ett urval på cirka $n = 350$ per månad.
- De är osäkra på om det finns periodicitet och väljer därför två kluster.
- Beräkna k och ge ett exempel på ett urval.

Referenser

ESS, 2013. Cluster sampling and multi-stage sampling.
URL http://essedunet.nsd.uib.no/cms/topics/weight/2/6.html

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

Skolverket, 2010. Attityder till skolan 2009: elevers och lärares attityder till skolan. Stockholm.