

# Surveymetodik

## Föreläsning 11

Måns Magnusson

Avd. Statistik, LiU

- 1 Klusterurval
  - Tvåstegs (eller flerstegs) klusterurval
- 2 Urval med olika inklusionssannolikheter
  - Vikter vid klusterurval

## Subsection 1

### Tvåstegs (eller flerstegs) klusterurval

- Första steget dras kluster/primära urvalsenheter (ex. skolor)
- I andra steget dras sekundära urvalsenheter (ex. elever)
- Det går att ha ytterligare steg om det finns behov
- Det kan vara olika urvalsförfarande i de olika stegen (ex. stratifierat urval och/eller OSU)  
(brukar kallas **komplexa surveyer**)

$N$  = antal psus (skolor) i populationen

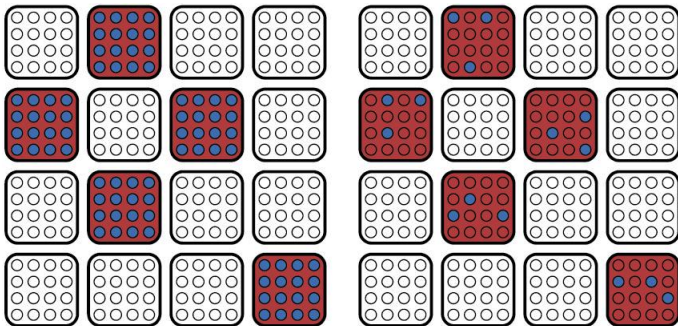
$n$  = antal psus (skolor) i urvalet

$M_i$  = antal ssus (elever) i skola  $i$

$m_i$  = antal ssus (elever) i psu (skola)  $i$  i vårt urval

$M_0$  = antal ssus (elever) i populationen

# Tvåstegs klusterurval - Exempel



Figur : Skillnad mellan enstegs och tvåstegs klusterurval. Källa: ESS (2013)

## Byggnaders energianvändning, tekniska status och inomhusmiljö (BETSI) Boverket (2009)

### Boverket

- **Syfte:** Kartlägga det svenska byggnadsbeståndet
- **Målpopulation:** Byggnader med taxeringsvärde på minst 50 tkr och med minst 50 m<sup>2</sup> samt individer i småbostadshus eller lägenhet
- **Urval:**
  - Flerstegsurval
    - Steg I: Stratifierat klusterurval av kommuner (pps/ $\pi$ ps)
    - Steg II: Stratifierat klusterurval av värderings/taxeringsenhet (OSU och pps/ $\pi$ ps)
    - Steg III: Klusterurval av byggnad (OSU)
    - Steg IV: Lägenhet (OSU)
- **Bortfall:** 21-35 % (beroende på byggnad)
- **Datainsamlingsmetod:** Besiktningar och pappersenkäter
- **Periodicitet:** Ett tillfälle (?)

- Vid enstegs klusterurval är

$$\hat{t}_{unb} = N\bar{t} = N \frac{1}{n} \sum_{i \in S} t_i$$

- Vid tvåstegs klusterurval är  $t_i$  inte känd, utan måste först skattas
- Vi kan i princip välja vilket skattningsmetod vi vill (unbiased, kvot- eller regressionsskattning)
- Vi skattar med den vanliga väntevärdesriktiga estimatoren

$$\hat{t}_i = M_i \bar{y}_i = M_i \sum_{j \in \mathcal{S}_i} \frac{y_{ij}}{m_i}$$

vilket ger (\*)

$$\hat{t}_{unb} = N\bar{t} = \frac{N}{n} \sum_{i \in S} M_i \bar{y}_i$$



- Vid enstegs klusterurval är

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

- Vid tvåstegs klusterurval måste vår osäkerhet i skattningen  $\hat{t}_i$  tas i beaktande
- Detta ger följande variansskattning

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

där

$$s_i^2 = \frac{1}{m_i - 1} \sum_{j \in S} (y_{ij} - \bar{y}_i)^2$$

- Variansen består av två delar - **inom** psu:s och **mellan** psu:s

- Precis som innan fungerar denna variansskattning bra om psu (ex. skolor) är ungefär lika stora
- Annars använder vi kvotskattning (se Lohr, 2009, s. 186)
- Den andra delen av variansskattningen (inom psu) är ofta betydligt mindre än den första delen (mellan psu)
- Därför används ibland

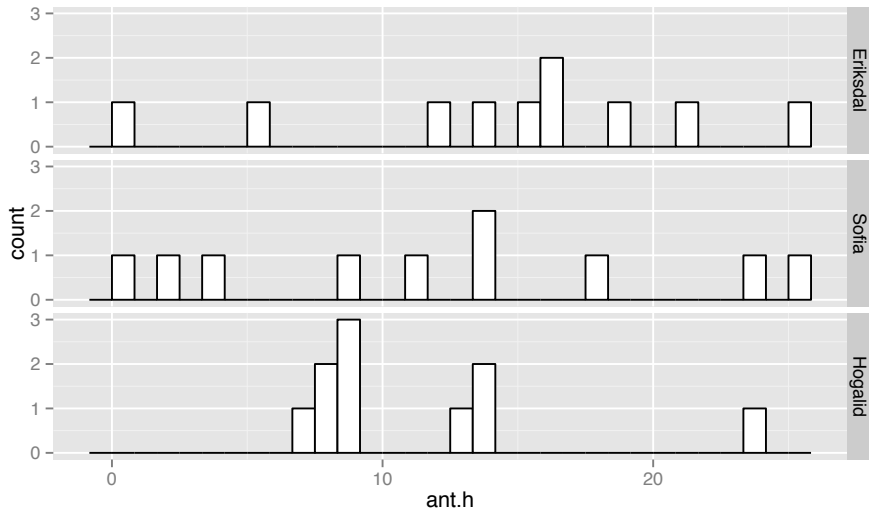
$$Var_{WR}(\hat{t}_{unb}) = N^2 \frac{s_t^2}{n}$$

- Konfidensintervallen beräknar vi (som vanligt)

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

- Vi vill undersöka IT-användningen i ett rektorsområde med 10 skolor.
- Syftet är att undersöka vid hur många lektioner datorer används i skolan i genomsnitt.
- Vi drar ett urval på 3 skolor och i varje skola drar vi 10 slumpvisa lärare.
- Vi vet att totalt arbetar 224 lärare i rektorsområdet.

# Exempel: IT-användning i skolan II



## Exempel: IT-användning i skolan III

Skola	$m_i$	$M_i$	$\bar{y}_i$	$\hat{t}_i$	$s_i$
Eriksdal	10	23	14.3	328.9	7.33
Sofia	10	19	12.1	229.9	8.66
Högalid	10	29	11.5	333.5	5.1

## Section 2

### Urval med olika inklusionssannolikheter

- Vid stratifiering såg vi att vi fick en lägre  $Var(\hat{t})$  om vi allokerade urvalet efter varians (Neymannallokering)
- Urval med olika inklusionssannolikheter - en "kontinuerlig stratifiering" efter en variabel  $x$
- **Probability proportional to size (pps)** vid urval med återläggning
- **$\pi$  proportional to size ( $\pi$ ps)** vid urval utan återläggning
- Ofta används pps/ $\pi$ ps med "storlek" som hjälpvariabel

## ■ Fördelar med pps/ $\pi$ ps

- Med pps/ $\pi$ ps får vi lägre  $Var(\hat{t})$  om  $x_i$  är korrelerad med  $Var(\hat{t}_i)$
- pps/ $\pi$ ps ger "självvägda" inklusionssannolikheter vid klusterurval - alla ssus (ex. elever) får samma inklusionssannolikheter

## ■ Nackdelar med pps/ $\pi$ ps

- Något mer komplicerad metod, särskilt  $\pi$ ps
- pps-urval är enklare matematiskt (både att dra urvalet och estimation) men risken finns att vi får dubletter (vid mindre  $N$ )



$\psi_i$  = sannolikhet för att dra kluster  $i$

$\mathcal{R}$  = Urvalsmängden (inklusive dubletter)

$\pi_i = P(\text{kluster } i \text{ ingår i urvalet}) = \text{inklusionssannolikhet för kluster } i$

$t_i$  = total i kluster  $i$

$\hat{t}_\psi$  = pps-skattning (Hansen-Hurwitz)

$\hat{t}_{HT}$  =  $\pi$ pps-skattning (Horwitz-Thompson)

- **Probability proportional to size** (pps) drar kluser med återläggning  
- en observation/kluster kan förekomma flera gånger
- Vi byter därför ut vår indikatorvariabel  $Z$  mot en antalsvariabel  $Q$
- Låt

$Q_i =$  antal gånger kluster *iförekommer* i urvalet

$$Q \sim \text{Bin}(n, \psi_i)$$

vilket ger att

$$E(Q_i) = n\psi_i = \pi_i$$

- För att skatta totalen använder vi oss av

$$\hat{t}_\psi = \frac{1}{n} \sum_{i \in \mathcal{R}} \frac{t_i}{\psi_i}$$

- $\hat{t}_\psi$  är väntevärdesriktig (\*)

- För att skatta variansen använder vi oss av

$$\hat{Var}(\hat{t}_\psi) = \frac{1}{n} \cdot \frac{\sum_{i \in \mathcal{R}} \left( \frac{t_i}{\psi_i} - \hat{t}_\psi \right)^2}{(n-1)}$$

- Eftersom det är ett urval med återläggning finns ingen ändlighetskorrektur.
- För att minimera  $Var(\hat{t}_\psi)$  så vill vi välja  $\psi_i$  proportionellt mot  $t_i$  - men vi känner inte  $t_i$  ...

- Ett vanligt sätt att välja  $\psi_i$  är att välja kluster proportionellt mot klusterstorleken (därav namnet pps)

$$\psi_i = \frac{M_i}{M_0} \text{ och } \sum_i^N \psi_i = 1$$

- Vikter kan beräknas på följande sätt

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

- Hur man drar urval (praktiskt) framgår i Lohr (2009, s. 225 ff.) (det är enkelt att göra i R med paketet `sampling`).

- Vi vill återigen skatta antalet lektioner då datorer används i undervisningen i ett givet rektorsområde med 10 skolor
- Vi vill nu dra ett pps-urval
- Denna gång undersöker vi samtliga lärare på skolan för en given vecka.

## Exempel: IT-användning i skolan II

	school	teachers	t_i	phi	t.hat	sample.pps	sample.srs
1	1	34	167	0.1604	1041	0	1
2	2	17	76	0.0802	948	0	0
3	3	27	124	0.1274	974	0	0
4	4	13	65	0.0613	1060	1	0
5	5	20	104	0.0943	1102	1	1
6	6	14	70	0.0660	1060	1	0
7	7	22	121	0.1038	1166	0	0
8	8	20	86	0.0943	912	0	0
9	9	21	110	0.0991	1110	0	0
10	10	24	132	0.1132	1166	0	1

- Den sanna totalen i populationen är 1055.

- Dra kluster med sannolikhet  $\psi_i$  (precis som vid enstegs klusterurval)
- Nu är  $t_i$  inte känd i varje urval eller kluster så  $t_i$  måste skattas
- Detta kan vi göra med vanligt OSU (eller vilken urvalsmetod som passar)
  - Det viktiga är dock att de olika klustertotalerna ( $t_i$ ) är oberoende av varandra
  - Om vi drar samma kluster flera gånger måste vi dra ett nytt OSU varje gång inom klustret

- Mer komplicerat än pps då urvalet inte är oberoende
  - När det första elementet är draget ändras sannolikheten att dra för de övriga
- Det finns flera olika metoder för att dra urval på ett sådant sätt att  $\pi_i$  är proportionellt efter en given variabel  
(de flesta finns implementerade i R-paketet `sampling`)
- Har vi dragit ett  $\pi$ ps-urval (eller ett vanligt OSU) kan vi använda Horwitz-Thompson-estimatorn (HT)

$$\hat{t}_{HT} = \sum_{i \in S} \frac{\hat{t}_i}{\pi_i}$$



- Det är enkelt att visa att HT-estimatorn är väntevärdesriktig (\*)
- Variansen är mer komplicerad

$$Var(\hat{t}_{HT}) = \sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} t_i^2 + \sum_{i=1}^N \sum_{k \neq i}^N \frac{\pi_{ik} - \pi_i \pi_k}{\pi_i \pi_k} t_i t_k$$

där  $\pi_{ik}$  är andra ordningens inklusionssannolikhet.

- Vid vanligt OSU är  $\pi_{ik}$  enklare att räkna ut

$$P(Z_i = 1 \text{ och } Z_k = 1) = \pi_{ik} = \frac{n}{N} \frac{n-1}{N-1}$$

- Men vid  $\pi$ ps-urval blir alla  $\pi_{ik}$  en matris mellan urvalsobjekten
  - Denna kan vara otymplig
  - I flera fall kan det vara så att den inte finns tillgänglig vid estimation
- Då kan variansestimatorn för pps-urval (med återläggning) användas istället

$$V_{WR}(\hat{t}_{HT}) = \frac{1}{n(n-1)} \sum_{i \in S} \left( \frac{t_i}{\pi_i} - \frac{\hat{t}_{HT}}{n} \right)^2$$

## Subsection 1

### Vikter vid klusterurval

- För att beräkna inklusionssannolikheten behöver vi

$$\pi_{ij} = P(j\text{:te enheten i kluster } i \text{ inkluderas i urvalet}) = \\ P(j\text{:te enheten} \mid \text{kluster } i) \cdot P(\text{kluster } i)$$

eftersom

$$P(A \cap B) = P(A \mid B) \cdot P(B)$$

- Detta ger inklusionssannolikheterna

$$\pi_{ij} = P(j\text{:te enheten} \mid \text{kluster } i) \cdot P(\text{kluster } i) = \frac{m_i}{M_i} \cdot \frac{n}{N}$$

och vikterna

$$w_{ij} = \frac{1}{\pi_{ij}} = \frac{M_i N}{m_i n}$$

vid tvästegsurval och

$$w_{ij} = \frac{1}{\pi_{ij}} = \frac{M_i N}{M_i n} = \frac{N}{n}$$

vid enstegsurval

- Med vikterna får vi följande skattning av  $\hat{t}_{unb}$

$$\hat{t}_{unb} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} \frac{M_i N}{m_i n} y_{ij} = \frac{N}{n} \sum_{i \in \mathcal{S}} M_i \bar{y}_i$$

och för

$$\hat{y}_r = \frac{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}}{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij}} = \frac{\frac{N}{n} \sum_{i \in \mathcal{S}} M_i y_{ij}}{\frac{N}{n} \sum_{i \in \mathcal{S}} m_i \frac{M_i}{m_i}} = \frac{\sum_{i \in \mathcal{S}} M_i y_{ij}}{\sum_{i \in \mathcal{S}} M_i}$$

Boverket, 2009. Statistiska urval och metoder i boverkets projekt betsi.  
Tech. rep., Boverket.

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.