Tecniche di indagine statistica

Lezione 24

Campionamento a due stadi

Motivazioni:
- elementi del cluster molto simili tra loro: spreco di risorse osservarli tutti
- molto costosa l'osservazione delle unità finali (ssu) rispetto a psu

Come si procede:
1. campione (usualmente CCS) di *m* unità di primo livello/stadio
2. campione (usualmente CCS) di unità di secondo livello/stadio entro le unità di primo stadio

Notazione come per grappoli:

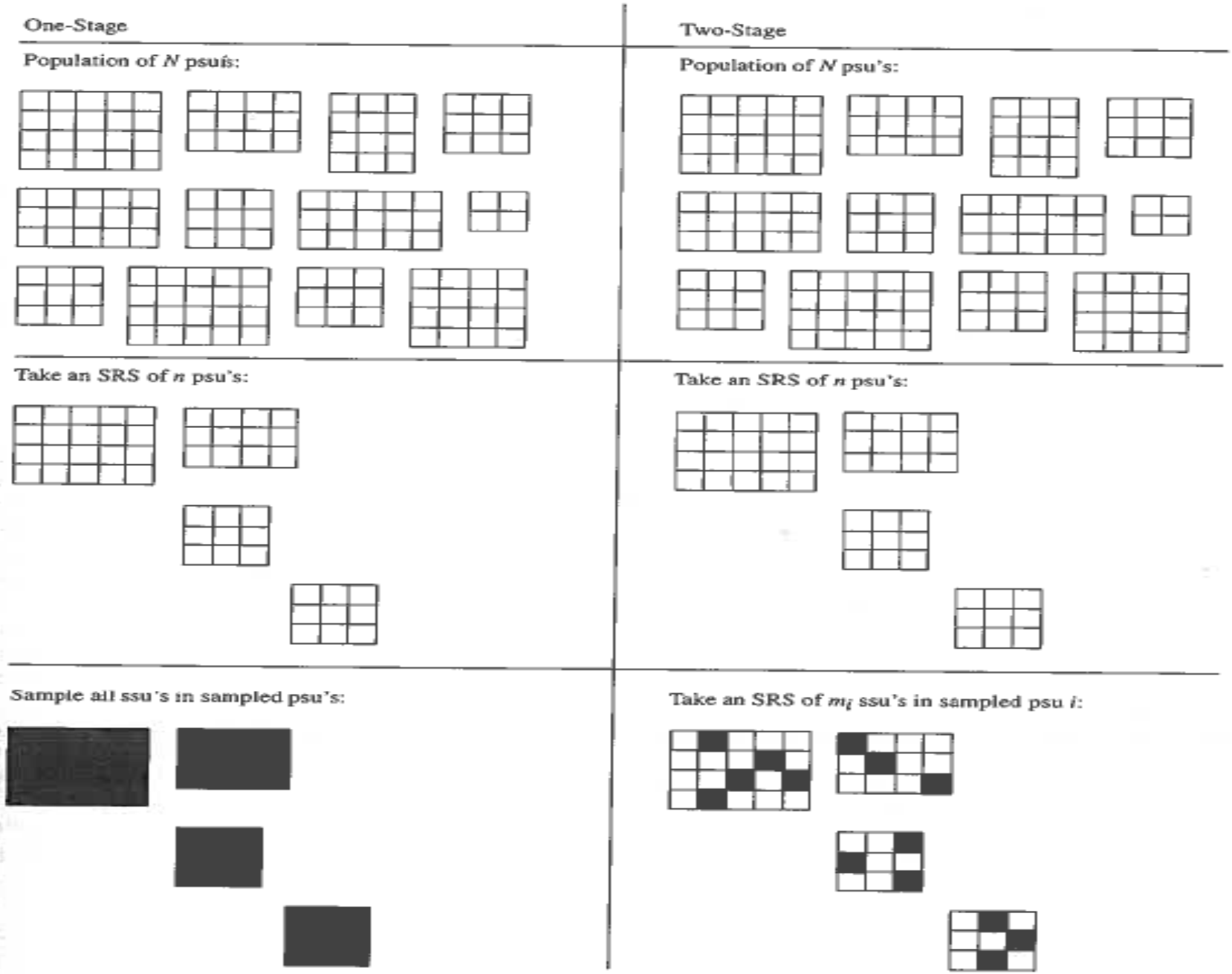
- campione $y_1^{(h)}, y_2^{(h)}, \ldots, y_{n_h}^{(h)}$ di dimensione $\boldsymbol{n_h}$ da psu h *(h = 1, 2, …, M)*

- totale nelle psu stimati da $y_T^{(1)}, y_T^{(2)}, \ldots, y_T^{(M)}$

- indicatore (0/1) psu nel campione $b_1, b_2, \ldots, b_M$

(selezione senza replicazione: 1 se selezionata, 0 se non selezionata;
se selezione con replicazione b_h = n.ro di volte psu *h* è nel campione (*h = 1, 2, …, M*)

Campionamento a grappoli *vs* due stadi



Campionamento a due stadi - stimatori

1. selezione di m unità di primo livello
2. selezione di n_h unità di secondo livello entro ogni unità selezionata al primo stadio

(selezione CCS in entrambi gli stadi: situazione più semplice)

Stimatore non distorto della media di Y nella pop.ne:

$$\bar{y}_{TS} = \frac{M}{N}\frac{1}{m}\sum_{h=1}^{M} b_h N_h \bar{y}^{(h)}$$

è la media degli stimatori per i totali nelle psu selezionate

come grappoli ma con stima totali psu

con varianza

$$var(\bar{y}_{TS}) = \left(\frac{M}{N}\right)^2 \left(1 - \frac{m}{M}\right)\frac{S_1^2}{m} + \frac{M}{mN^2}\sum_{h=1}^{M} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{S_{2,h}^2}{n_h}$$

Dove $S_1^2 = \frac{1}{M-1}\sum_{h=1}^{M}\left(Y_T^{(h)} - \bar{Y}_T\right)^2$ varianza dei totali delle psu

$S_{2,h}^2 = \frac{1}{N_h-1}\sum_{k=1}^{N_h}\left(Y_k^{(h)} - \bar{Y}^{(h)}\right)^2$ varianza entro psu h

(stimate dalle quantità campionarie s_1^2 e $s_{2,h}^2$). Se N grande, 2^ termine trascurabile

Scelte per formare un campione su più stadi

Probabilità di inclusione delle unità – due stadi

Prob osservare nel campione la ssu k che appartiene alla psu h =

$$p_{hk} = p_h p_{k/h} = \frac{m}{M}\frac{n_h}{N_h}$$

ogni unità selezionata rappresenta sé stessa e $MN_h/(mn_h - 1)$ altre unità (in totale = MN_h/mn_h)

Per avere campione auto-ponderante:
n_h proporzionale a N_h così n_h/N_h è ≈ costante in tutte le psu

Per avere campione:

1. auto-ponderante
2. ridurre impatto omogeneità intraclasse
3. garantire dimensione campionaria = *n* (fissato) indipendentemente dalle psu selezionate al primo stadio

selezione psu con replicazione e probabilità proporzionale a N_h (probabilità variabile) e selezione numero fisso di ssu in ogni psu

⇒ campione a due stadi PPS $\quad p_{hk} = \frac{mN_h}{M}\frac{n_0}{N_h}$

ampiezza campionaria
$n = mn_0$

Numerosità ottima ssu (n_o) in funzione dei costi

Modello di costo $C = mC_h + mn_0c$

C costo totale
C_h costo per psu
c costo per elemento (ssu)

$n_0 \approx \sqrt{\frac{C_h}{c}\frac{(1-\rho)}{\rho}}$ Poiché $campione\ n = mn_0$, fissato n e n_0, poi si determina m

a parità di altre condizioni, più l'omogeneità interna è elevata, più alti i costi per unità (c) e più bassi i costi per gruppo (C_h)

⇩

più il campione sarà sparpagliato tra i gruppi (psu)

N.B. indagini multiscopo: usuali considerazioni

Campionamento con probabilità variabili PPS
(Probability Proportional to Size)

In generale, con probabilità variabili di selezione per le psu:

- è favorita in termini probabilistici l'entrata nel campione delle unità di grandi dimensioni (campione "rappresenta" meglio la popolazione rispetto a uno selezionato con probabilità uguali)
- le unità finali sono estratte da blocchi mediamente più estesi, e quindi sono più disperse e la stima è più efficiente di un campione selezionato con probabilità costanti ad ogni stadio

Spesso, selezione di una unica psu ($m = 1$):

- se stratificazione al primo stadio, ogni strato può contenere poche psu
- possono anche essere definiti un grande numero di strati per aumentare la precisione

Ovviamente, con una psu non è possibile ottenere stime della variabilità tra psu entro lo strato: procedure specifiche per stimare la varianza

Come si fa la selezione PPS?

Campionamento con probabilità variabili PPS
caso semplice: selezione con replicazione

P (unità h è selezionata alla prima estrazione) = ψ_h
= *P* (unità h è selezionata alla seconda estrazione) = *P* (terza) …

Idea sottostante:
- selezione di m psu con replicazione
- stimare il totale per ciascuna psu
- se psu replicate, il totale sarà incluso tante volte quante la psu è stata selezionata
- **stima** totale popolazione = media delle m stime t_h indipendenti
- **stima varianza** = varianza campionaria delle m stime indipendenti diviso m (stima della componente S_1^2)

Per *selezione PPS* (conoscenza di una misura di dimensione **per tutte** le psu nella popolazione):

1. Metodo della cumulata
2. Metodo di Lahiri (particolarmente utile quando il n.ro di psu è grande)

Metodo della cumulata per PPS

647 studenti in 15 classi, campione di 5 classi con replicazione e prob. proporzionale a N_h (= n.ro studenti per classe)

Class Number	N_h	ψ_h	Cumulative N_h Range	
1	44	0.068006	1	44
2	33	0.051005	45	77
3	26	0.040185	78	103
4	22	0.034003	104	125
5	76	0.117465	126	201
6	63	0.097372	202	264
7	20	0.030912	265	284
8	44	0.068006	285	328
9	54	0.083462	329	382
10	34	0.052550	383	416
11	46	0.071097	417	462
12	24	0.037094	463	486
13	46	0.071097	487	532
14	100	0.154560	533	632
15	15	0.023184	633	647
Total	647	1		

$$\psi_h = N_h/647$$

1. Generazione di 5 numeri casuali : 487, 369, 221, 326, 282
2. Classi nel campione: 13, 9, 6, 8, 7

(se n.c.= 553, 082, 245, 594, 150, campione: 14, 3, 6, 14, 5 con classe 14 inserita 2 volte)

Si utilizza anche selezione sistematica (che produce campioni non replicati ma in grandi pop.ni, differenza minima)

Selezione sistematica per PPS /1

(che produce campioni non replicati ma in grandi pop.ni, risultati molto simili)

647 studenti in 15 classi, campione di 5 classi:

- lista degli elementi per la prima psu, poi la seconda e così via
- selezione sistematica dalla lista

1 < *x* < *129* (647/5 ≈ 129.4), psu nel campione: *x*, *x*+*129*, …

se *x* = 112

Number in Systematic Sample	psu Chosen
112	4
241	6
370	9
499	13
628	14

N.B.:

Non vero campione con replicazione, poiché classi <= 129 non entrano più di una volta nel campione e classi > 129 hanno $P = 1$ di far parte del campione

ma facile da fare !

(se psu organizzate geograficamente, campione ottenuto è più sparso con risultati migliori)

Metodo di Lahiri (*rejective method*) per PPS

M = n.ro psu, *max* (N_h) = dimensione massima psu

1. selezione numero casuale (n.c.) tra 1 e M (psu da considere)

2. selezione n.c. tra 1 e *max* (N_h):

   - n.c. <= N_h psu h è inclusa nel campione
   - altrimenti si torna al punto 1

3. ripetere fino a ottenere il numero di psu (ampiezza campionaria 1^ stadio) desiderato.

Esempio 15 classi: *max* (studenti) = 100, generazioni di coppie di n.c.:

step 1: n.c. da 1 a 15

step 2: n.c. da 1 a 100

Metodo di Lahiri - esempio

15 classi: *max* (studenti) = 100, generazioni di coppie di n.c.,

1^: 1, …,15 (psu); 2^: 1,…,100 (per decidere se tenere psu)

primo n.c. (psu h)	secondo n.c.	N_h	Decisione
12	6	24	$6 < 24$; include psu 12 in sample
14	24	100	Include in sample
1	65	44	$65 > 44$; discard pair of numbers and try again
7	84	20	$84 > 20$; try again
10	49	34	Try again
14	47	100	Include
15	43	15	Try again
5	24	76	Include
11	87	46	Try again
1	36	44	Include

Confronto selezione non probabilistica vs probabilistica

Online (access) panel
Canale televisivo privato.
Campione di **3000** rispondenti

Canale televisivo.
Campione probabilistico
Random Digit Dialing
di 1200 rispondenti

Table 11.2 Dutch Parliamentary Elections 2003: Outcomes and the Results of Various Opinion Surveys

	Election	Kennisnet	RTL4	SBS6	Nederland 1
Sample size		17.000	10.000	3.000	1.200
Seats in Parliament					
CDA (Christian democrats)	44	29	24	42	42
LPF (populist party)	8	18	12	6	7
VVD (liberals)	28	24	38	28	28
PvdA (social democrats)	42	13	41	45	43
SP (socialists)	9	22	10	11	9
GL (green party)	8	26	9	6	8
D66 (liberal democrats)	6	4	7	5	6
Other parties	5	14	9	7	7
Mean absolute difference		12.5	5.3	1.8	0.8

Sito Web su temi legati all' istruzione (partecipavano 11.000 scuole e altre istituzioni educative ma anche altri)
Circa **17.000** rispondenti

Canale televisivo pubblico
Circa **10.000** rispondenti

Confronto selezione non probabilistica vs probabilistica

Panel online (pesati)

Table 11.3 Dutch Parliamentary Elections 2006: Outcomes and the Results of Various Opinion Surveys

	Election Result	Politieke Barometer	Peil.nl	De Stemming	DPES 2006
Sample size		1000	2500	2000	2600
Seats in Parliament					
CDA (Christian democrats)	41	41	42	41	41
PvdA (social democrats)	33	37	38	31	32
VVD (liberals)	22	23	22	21	22
SP (socialists)	25	23	23	32	26
GL (green party)	7	7	8	5	7
D66 (liberal democrats)	3	3	2	1	3
ChristenUnie (Christian)	6	6	6	8	6
SGP (Christian)	2	2	2	1	2
PvdD (animal party)	2	2	1	2	2
PvdV (conservative)	9	4	5	6	8
Other parties	0	2	1	2	1
Mean absolute difference		1.27	1.45	2.00	0.36

Dutch Parliamentary Election Study - campione probabilistico a 2 stadi (CAPI). Statistics Netherlands

Testo *Survey Methods and Practices*, Statistics Canada

Capitolo 6 'Sample designs'

in particolare

- 6.2.5
- 6.2.7

Capitolo 7 'Estimation'

in particolare

- 7.1.1
- 7.1.3 (e 7.1.4.1)
- 7.2 (no 7.2.3)
- 7.3 (se curiosi 7.3.4)

Capitolo 8 'Sample Size Determination and Allocation'

dare un'occhiata