Tecniche di indagine statistica

Lezione 23

Campionamento a grappoli/stadi

solo i *clusters* (unità di primo stadio) selezionati al primo stadio devono rappresentare tutta la popolazione
i gruppi (clusters) dovrebbero essere quindi *molto eterogenei* al loro interno

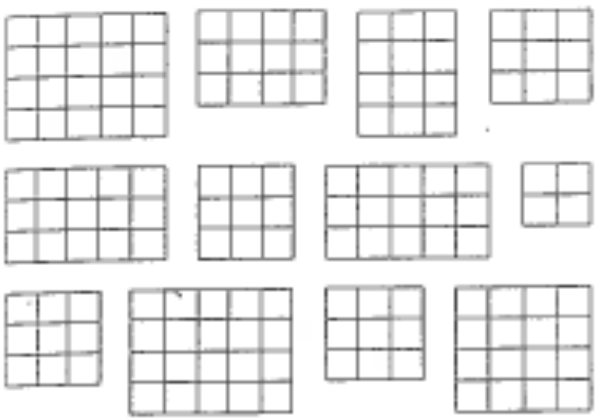
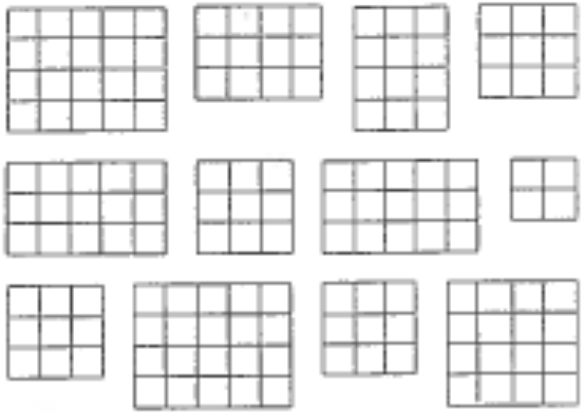
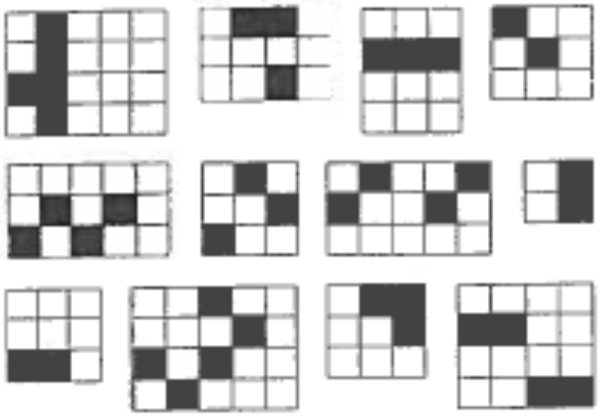
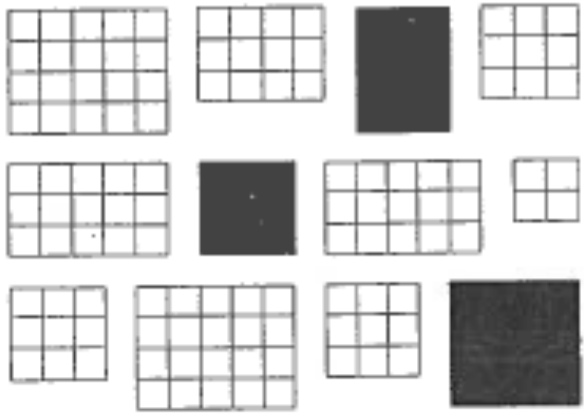
in realtà

l'appartenenza ad un gruppo fa sì che le unità risultino interdipendenti o omogenee o correlate tra loro (a causa di fattori misurabili e non: *condivisione di uno stesso contesto/ esperienze simili*)
le informazioni "originali" sono perciò "inferiori" al numero di unità del gruppo (selezionando tutte le unità del cluster, si ripete parzialmente una informazione già nota)

⇨ stime meno efficienti

usato poiché meno costoso e molto conveniente dal punto di vista operativo selezionare clusters che non selezionare casualmente dalla popolazione

Campionamento stratificato *vs* grappoli

Stratified Sampling	Cluster Sampling
Each element of the population is in exactly one stratum.	Each element of the population is in exactly one cluster.
Population of H strata: stratum h has n_h elements: 	One-stage cluster sampling: population of N clusters: 
Take an SRS from *every* stratum: 	Take an SRS of clusters; observe all elements within the clusters in the sample: 
Variance of the estimate of $\bar{y}_U$ depends on the variability of values *within* strata.	The cluster is the sampling unit: the more clusters we sample, the smaller the variance. The variance of the estimate of $\bar{y}_U$ depends primarily on the variability *between* cluster means.
For greatest precision, individual elements within each stratum should have similar values, but stratum means should differ from each other as much as possible.	For greatest precision, individual elements within each cluster should be heterogeneous, and cluster means should be similar to one another.

in ogni strato è selezionato un campione di unità

nel campione di *strati*, tutti gli elementi sono osservati

Notazione - Campionamento a grappoli

(scuole sup. Fvg = 140 a.s. 2010/11, = 189 a.s. 2020/21)

U = popolazione suddivisa in $\boldsymbol{M}$ gruppi/clusters (psu/up) di numerosità N_h $(h = 1, 2, \ldots, M)$ ciascuno, con $\sum_{h=1}^{M} N_h = N$

(studenti nella scuola)

(Fvg a.s. 2010/11: 46077, 2020/21: 49516)

N_h valori di Y nel gruppo h: $Y_1^{(h)}, Y_2^{(h)}, \ldots, Y_{N_h}^{(h)}$

(Y= essere ripetente), $t^{(h)}$ = n.ro ripetenti

Per ciascun cluster:

Media di Y nel cluster h: $\bar{Y}^{(h)} = \frac{1}{N_h} \sum_{k=1}^{N_h} Y_k^{(h)}$

Totale di Y nel cluster h: $Y_T^{(h)} = \sum_{k=1}^{N_h} Y_k^{(h)} = N_h \bar{Y}^{(h)}$

Popolazione:

(ripetenti a.s. 2010/11: 3041, a.s. 2020/21: 297)

Media di Y: $\bar{Y} = \frac{1}{N} \sum_{h=1}^{M} N_h \bar{Y}^{(h)}$

(% ripetenti Fvg 2010/11: 6.6%, a.s. 2020/21: 0.6%

Totale medio: $\bar{Y}_T = \frac{1}{M} \sum_{h=1}^{M} Y_T^{(h)} = \frac{1}{M} \sum_{h=1}^{M} N_h \bar{Y}^{(h)}$

"No matter how you define it, the notation for cluster sampling is messy, because you need notation for both psu and the ssu levels." (Lohr, 2022)

Campionamento a grappoli – con probabilità uguali

Campione CCS di m clusters (con uguale probabilità)

Totali di Y nei clusters selezionati: $y_T^{(1)}, y_T^{(2)}, \ldots, y_T^{(m)}$

Stimatore non distorto del totale medio: $\bar{y}_T = \frac{1}{m}\sum_{i=1}^{m} y_T^{(i)}$

Stimatore non distorto della media di Y nella pop.ne: $\bar{y}_{CL} = \frac{M}{N}\bar{y}_T$

con varianza

$$var(\bar{y}_{CL}) = \left(\frac{M}{N}\right)^2 \frac{1-f}{m} S_C^2 \text{ , } f = m/M \text{ e } S_C^2 = \frac{1}{M-1}\sum_{h=1}^{M}\left(Y_T^{(h)} - \bar{Y}_T\right)^2$$

stimata da $s_C^2 = \frac{1}{m-1}\sum_{j=1}^{m}\left(y_T^{(h)} - \bar{y}_T\right)^2$

varianza determinata da differenze tra i totali di clusters e NON entro i clusters.

Camp.to grappoli: usato in molte indagini in cui il costo di campionamento per unità è trascurabile rispetto al costo di campionamento del cluster

(classe scolastica/scuola: cluster *naturale* per indagini su istruzione. Intervistare tutti gli studenti in una classe aumenta di poco i costi rispetto ad intervistarne solo alcuni)

- Campione auto-ponderante $w_j^{(h)} = M/m$

Campionamento a grappoli – con probabilità uguali

Campione a grappoli equivalente a CCS a livello aggregato:

Le unità selezionate sono i clusters e le osservazioni sono i totali dei valori $Y_j^{(h)}$ delle unità nei clusters selezionati

$(j = 1, \dots, N_h$ e $h = 1, \dots, M)$

Selezione clusters con probabilità uguali:

1. varianza stimatore data da variazioni nei totali di clusters $Y_T^{(h)}$.
   Se valori individuali $Y_i^{(h)}$ non molto variabili, la variazione nei totali data principalmente dal numero di unità nel cluster N_h.
2. numero di unità individuali da osservare nel campione può essere molto variabile se N_h molto diversi tra loro
   (es: va bene per selezione di classi scolastiche, che più o meno hanno la stessa dimensione $N_h = L$ ma non con selezione scuole)
3. se ampiezza clusters diverse, più conveniente selezionare i clusters con probabilità variabile (proporzionale all'ampiezza)

Confronto CCS e camp.to a grappoli ($N_h = L$)

Camp.to a grappoli: sempre stimatori meno precisi di CCS di pari numerosità

Situazione opposta a campionamento stratificato
(aumenta precisione se var *tra* gli strati grande rispetto a var *entro*)

poiché in camp.to a grappoli variabilità stimatore dipende interamente dalla variabilità *tra* i clusters
(diminuisce precisione se var *tra* i clusters è grande)

Spesso, elementi in clusters diversi più variabili che elementi nello stesso cluster, poiché clusters diversi hanno totali/medie diverse (es. diverso rendimento di classi di studenti, dovuto a insegnanti/contesti diversi)

Quanto "simili" sono tra loro gli elementi di un cluster?

Misura del grado di omogeneità interna ai clusters

Quanto "simili" sono tra loro gli elementi di un cluster?

Coefficiente di correlazione *intraclasse* (con uguale ampiezza $N_h = L$):

$$\rho = 2\sum_{h=1}^{M}\sum_{j<k}\left(\left(Y_j^h - \bar{Y}\right)\left(Y_k^h - \bar{Y}\right)\right)\Big/\left[(L-1)(ML-1)S^2\right]$$

$$Deff = \frac{var(\bar{y}_{cL})}{var(\bar{y}_{\mathrm{CCS}})} \approx \frac{ML-1}{L(M-1)}\left[1 + (L-1)\rho\right]$$

se M grande:

$ML - 1 \approx L(M-1)$, il rapporto è $\approx [1 + (L-1)\rho]$

con $\rho = 0.5$ e $L = 5$,

deff $\approx [1 + (L-1)\rho] = 3$

Deff = 3 indica che servono 300 elementi con campione a grappoli per ottenere la precisione di 100 elementi in CCS
(in cluster "naturali" $\rho > 0$)

Esempio campione a grappoli stessa dimensione

Stima GPA medio degli studenti in casa studente:
400 (*N*) studenti in 100 (*M*) appartamenti (suites) da 4 (*L*)

1. Per CCS serve lista studenti per selezionare *n* studenti

2. Campione a grappoli: selezione CCS di *m*=5 suites e chiesto GPA ai 4 studenti delle 5 suites (osservazioni totali = 20):

	App.to					
Studente	1	2	3	4	5	
1	3.08	2.36	2.00	3.00	2.68	
2	2.60	3.04	2.56	2.88	1.92	
3	3.44	3.28	2.52	3.44	3.28	
4	3.04	2.68	1.88	3.64	3.20	
Totale	12.16	11.36	8.96	12.96	11.08	$y_T^{(m)}$

$\bar{y}_T = \frac{1}{m}\sum_{i=1}^{m} y_T^{(i)}$ = (12.16+11.36+8.96+12.96+11.08)/5 = 11.304

$\bar{y}_{CL} = \frac{M}{N}\bar{y}_T$ = (100 * 11.304)/400 = 2.826

Esempio campione a grappoli stessa dimensione

Stima varianza

	App.to				
Studente	1	2	3	4	5
1	3.08	2.36	2.00	3.00	2.68
2	2.60	3.04	2.56	2.88	1.92
3	3.44	3.28	2.52	3.44	3.28
4	3.04	2.68	1.88	3.64	3.20
Totale	12.16	11.36	8.96	12.96	11.08

$$\widehat{var}(\bar{y}_{CL}) = \left(\frac{M}{N}\right)^2 \frac{1-f}{m} s_C^2 =$$

$$s_C^2 = \frac{1}{m-1}\sum_{j=1}^{m}\left(y_T^{(h)} - \bar{y}_T\right)^2$$

$$= \frac{1}{5-1}\left[(12.16 - 11.304)^2 + \ \ldots + (11.08 - 11.304)^2\right] = 2.256$$

$$\widehat{var}(\bar{y}_{CL}) = \left(\frac{100}{400}\right)^2 \frac{1}{5}\left(1 - \frac{5}{100}\right) 2.256 = 0.02679$$