Tecniche di indagine statistica

Lezione 19 e 20

Quando si può fare di meglio del CCS ?

- Si hanno informazioni *ausiliarie* sulla popolazione

(variabile *X* – correlata con *Y* - di cui è nota una misura di sintesi a livello di popolazione [t_X, M_X]. *X* può essere rilevata per le *n* unità del CCS e si usa lo *stimatore rapporto* per stima parametro *Y*:

$$\hat{t}_{Y_R} = \hat{B}\, t_X \qquad \hat{B} = {}^{\hat{t}_Y}/_{\hat{t}_X} \qquad \hat{t}_{Y_R} = \hat{t}_Y\, {}^{t_X}/_{\hat{t}_X}$$

in genere non corretti ma più precisi)

N.b.: campione ancora CCS!

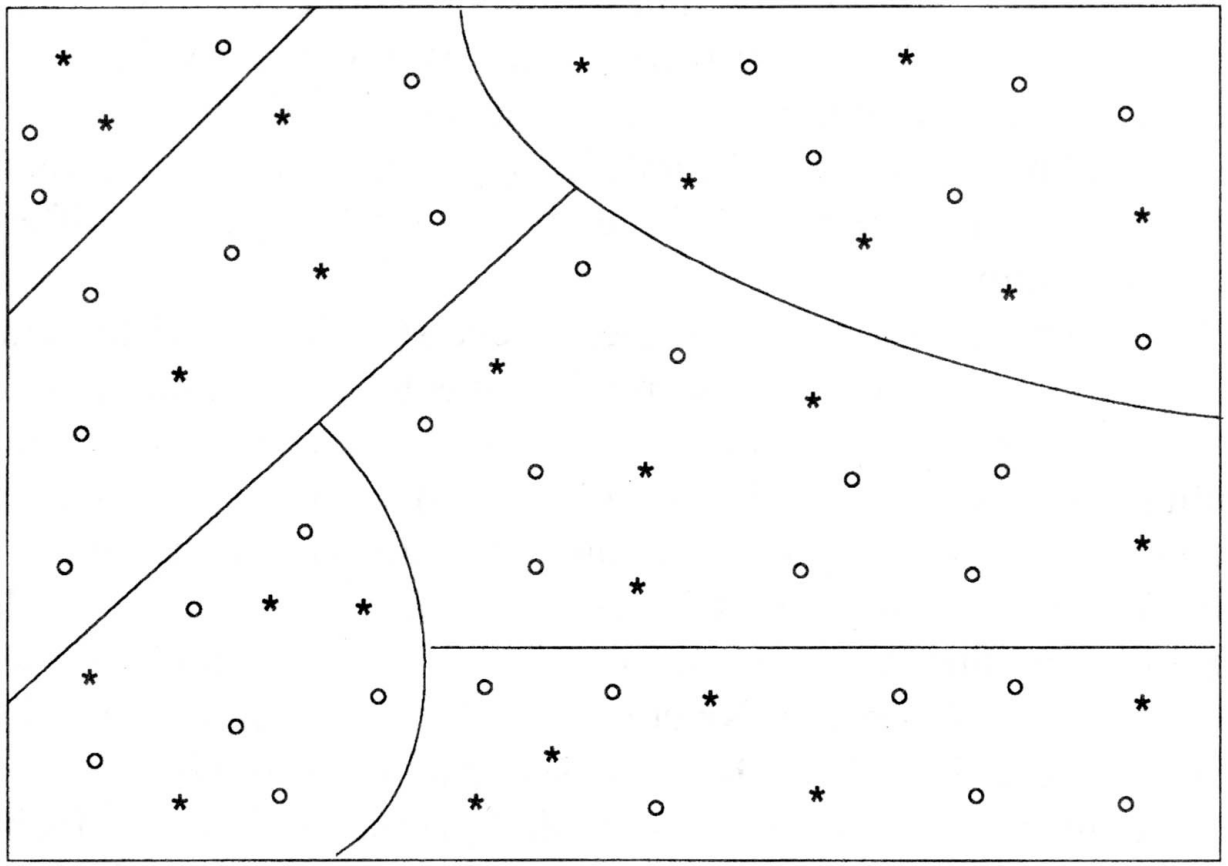
- La popolazione è suddividibile in *gruppi omogenei* al loro interno.

- I costi per raggiungere le unità possono *variare notevolmente* e disegni diversi comportano costi molto inferiori.

Campione stratificato

popolazione è suddividibile in gruppi omogenei al loro interno

Spesso, informazioni (supplementari) su caratteristiche (X_i) delle unità sono note a livello di popolazione e possono essere usate *per progettare il campione*



Tali caratteristiche identificano gruppi/ sottopopolazioni (strati) della popolazione che conviene considerare nel campione se la variabile Y presenta valori medi diversi nei vari strati

Fig. 2 Campionamento stratificato: unità della popolazione (◦), unità campionaria (⋆)

Campione stratificato - Esempio

popolazione è suddividibile in gruppi omogenei al loro interno

Età (*X*) e reddito mensile (*Y*, in euro) di N = 9 soggetti

n. soggetto	1	2	3	4	5	6	7	8	9
età	30 anni	30 anni	30 anni	40 anni	40 anni	40 anni	50 anni	50 anni	50 anni
reddito	2000	2100	2200	3000	3100	3300	4000	4100	4200

Media e deviazione standard totale e per i tre gruppi di età

	30 anni	40 anni	50 anni	Totale
Media	2100	3100	4100	3100
Dev.standard	81,65	81,65	81,65	820,57

- Variabilità nella popolazione superiore a quella nei singoli strati per età
- Variabilità totale (pop.ne): risultato di *variabilità* entro i singoli strati e *variabilità* tra gli strati

Campione stratificato - Esempio

popolazione è suddividibile in gruppi omogenei al loro interno

Per selezione di un campione $n = 3$, due alternative:

1. CCS con $n = 3$ da pop.ne di 9 unità con selezione casuale o sistematica

2. Selezione di 3 CCS indipendenti con $n_h = 1$ ($h = 1,\ldots, 3$) entro ciascun strato di età (campione totale formato dalle 3 unità selezionate in ciascun strato)

Quale alternativa è preferibile?

Quali caratteristiche dello stimatore nel caso 2?

N.b.: Esempio ad hoc per introdurre la logica del campionamento stratificato!!!

Campione stratificato - motivazioni

Incorpora nel disegno informazioni sulla popolazione
e consente/richiede:

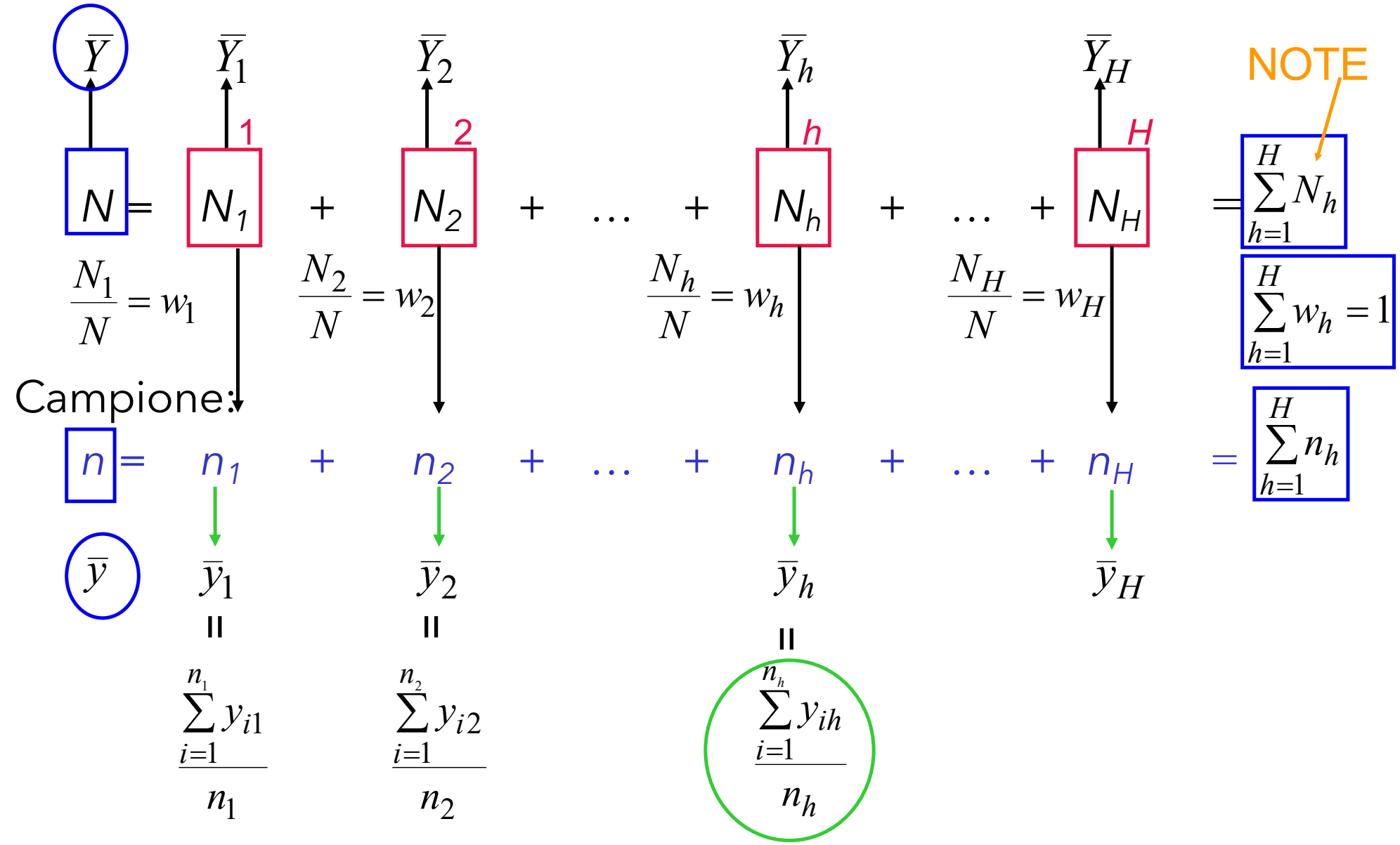
1. "garanzia" contro campioni che, per puro effetto del caso, potrebbero sembrare "poco" rappresentativi
2. stime per gruppi separati (con precisione comparabile)
3. liste disponibili per aree/gruppi separati
4. uso di differenti modalità (costruzione lista, rilevazione, campionamento) nei vari strati per praticità e/o riduzione costi
5. migliora l'efficienza delle stime (maggiore precisione rispetto CCS poiché la varianza entro gli strati è, spesso, più contenuta della varianza nell'intera popolazione)

Campione stratificato – procedura

1. La popolazione di N unità è classificata in H strati a seconda di informazioni supplementari
   (es. sesso, età, corso di studio, assets bancari, …)
   Classificazione negli strati deve essere nota prima della selezione (*ogni unità appartiene a un solo strato*)
2. E' selezionato un campione di numerosità $n_h, h = 1, \dots, H$ dalle N_h unità di ogni strato:
   a. se $N_h > 1$, almeno $n_h = 2$ per avere stime della variabilità nello strato
   b. campione negli strati: in genere CCS mediante procedure di selezione *casuale* o *sistematica*
3. Numerosità totale del campione (fissata prima) $n = \sum_{h=1}^{H} n_h$

Campione Stratificato – notazione generale

Popolazione:



Campione Stratificato – stimatore media pop.ne

Media

Popolazione $\bar{Y} = \sum_{h=1}^{H} w_h \bar{Y}_h$ con (noti) $w_h = \frac{N_h}{N}$

Media campionaria $\bar{y}_{str} = \sum_{h=1}^{H} w_h \bar{y}_h$

$$= \sum_{h=1}^{H} \frac{N_h}{N} \left(\sum_{i=1}^{n_h} \frac{y_{ih}}{n_h} \right)$$

Varianza campionaria

$$\hat{\text{var}}(\bar{y}_{str}) = \sum_{h=1}^{H} w_h^2 \underbrace{\hat{\text{var}}(\bar{y}_h)}_{\text{se CCS entro strati}}$$

$$= (1 - f_h) \frac{s_h^2}{n_h} \qquad s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$$

Intervallo di confidenza $\bar{y}_{str} \pm z_{\alpha/2} \hat{SE}(\bar{y}_{str})$

(con:
1. n_h grande e
2. tanti strati, Krewski e Rao, 1981)

Campione Stratificato – stimatore del totale

Totale

Popolazione

$$t = \sum_{h=1}^{H} t_h \qquad t_h = \sum_{i=1}^{N_h} y_{ih}$$

Totale campionario

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

Varianza campionaria

$$var(\hat{t}_{str}) = \sum_{h=1}^{H} V(\hat{t}_h) = \sum_{h=1}^{H} (1 - \frac{n_h}{N_h}) N_h^2 \frac{S_h^2}{n_h}$$

$$\widehat{var(\hat{t}_{str})} = \sum_{h=1}^{H} (1 - \frac{n_h}{N_h}) N_h^2 \frac{s_h^2}{n_h}$$

Campione Stratificato – stimatore proporzione

Proporzione campionaria

$$\hat{p} = \sum_{h=1}^{H} w_h \hat{p}_h \qquad \text{con} \qquad \hat{p}_h = \sum_{i=1}^{n_h} \frac{n^*_{hi}}{n_h}$$

$$\widehat{\text{var}}(\hat{p}) = \sum_{h=1}^{H} w_h{}^2 \hat{p}_h \frac{(1-\hat{p}_h)}{n_h - 1}(1-f_h)$$

n^*_{hi} indica che l'unità *i* nello strato *h* presenta la caratteristica di interesse

Allocazione di *n* entro gli strati – allocazione proporzionale

Frazione di campionamento negli *H* strati uguale a quella totale:

$$f_h = f \Rightarrow \frac{n_h}{N_h} = \frac{n}{N} \Rightarrow n_h = w_h n$$

(la probabilità di inclusione $\pi_{ih} = n_h / N_h$ è uguale per ogni unità in ogni strato)

$$\bar{y}_p = \sum_{h=1}^{H} w_h \left(\sum_{i=1}^{n_h} y_{hi} \Big/ n_h \right) = \sum_{h=1}^{H} \frac{N_h}{N} \left(\frac{\sum_{i=1}^{n_h} y_{hi}}{n_h} \right) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{y_{hi}}{n} = \sum_{i=1}^{n} \frac{y_i}{n}$$

$$\frac{N}{n}$$

$$\widehat{\mathrm{var}}(\bar{y}_p) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 (1 - f_h) \frac{s_h^2}{n_h} = (1-f) \sum_{h=1}^{H} \frac{N_h}{N \cdot N} s_h^2 \frac{N_h}{n_h} = \frac{(1-f)}{n} \sum_{h=1}^{H} w_h s_h^2$$

$$\hat{p}_p = \sum_{h=1}^{H} w_h \hat{p}_h = \sum_{i=1}^{n} \frac{n_i^*}{n}$$

$$\widehat{\mathrm{var}}(\hat{p}_p) = \frac{(1-f)}{n^2} \sum_{h=1}^{H} \frac{n_h^2}{n_h - 1} \hat{p}_h (1 - \hat{p}_h)$$

Varianza *media* di strato

Varianza stimatore media campionaria e totale in campione stratificato proporzionale

$$f_h = f = \frac{n_h}{N_h} = \frac{n}{N} \qquad w_h = \frac{N_h}{N} \qquad \Rightarrow \quad n_h = \frac{n}{N} N_h = n w_h$$

Media campionaria

$$\widehat{var}(\bar{y}_p) = \sum_{h=1}^{H} w_h^2 \, \widehat{var}(\bar{y}_h)$$

$$= \sum_{h=1}^{H} {w_h}^2 (1 - f_h) \frac{{s_h}^2}{n_h}$$

$$= \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 (1 - f_h) \frac{{s_h}^2}{n_h}$$

$$= (1 - f) \sum_{h=1}^{H} \frac{N_h}{N^2} {s_h}^2 \frac{N_h}{n_h} = \frac{N}{n} \quad = \frac{(1 - f)}{n} \sum_{h=1}^{H} w_h {s_h}^2$$

Totale

$$\widehat{var}(\hat{t}_p) = \sum_{h=1}^{H} (1 - \frac{n_h}{N_h}) N_h N_h \frac{s_h^2}{n_h} = (1 - \frac{n}{N}) \frac{N}{n} \sum_{h=1}^{H} N_h \, s_h^2$$

Stratificazione e efficienza delle stime

Design effect $Deff = \frac{\text{var}(z)}{\text{var}(z_{cs})} \geq 1$

z stimatore disegno complesso

z_{cs} stimatore campione casuale semplice

con campione stratificato, in generale:

$$Deff(\bar{y}_{st}) = \frac{\text{var}(\bar{y}_{st})}{\text{var}(\bar{y}_{cs})} \leq 1$$

poiché:

$$var(\bar{y}_{cs}) \cong \underbrace{var(\bar{y}_{st})}_{\text{entro gli strati}} + \frac{(1-f)}{n} \underbrace{\sum w_h (\bar{y}_h - \bar{y})^2}_{\text{tra gli strati} \geq 0} \Longrightarrow \text{var}(\bar{y}_{cs}) \geq \text{var}(\bar{y}_{st})$$

Il guadagno è maggiore, data la variabilità S^2 nella popolazione, quanto più gli strati sono eterogenei tra loro e omogenei al loro interno

(con proporzioni il guadagno è più modesto)

Omogeneità interna: strati di grande dimensione, in genere, sono più variabili di strati piccoli

Stratificato proporzionale e *deff* – esempio 1

Campione di 300 studenti da una popolazione di 3000 per stima proporzione

$$f = \frac{1}{10} = \frac{300}{3000}$$

Facoltà	N_h	W_h	n_h	c_h	p_h	$n_h p_h(1-p_h)$
Economia	950	0,32	95	86	0,905	8,147
Sociologia	430	0,14	43	22	0,512	10,744
Statistica	250	0,08	25	18	0,720	5,040
Sc. Politiche	390	0,13	39	31	0,795	6,359
Giurisprudenza	320	0,11	32	20	0,625	7,500
Storia	660	0,22	66	33	0,500	16,500
	3000	**1**	**300**	**210**		**54,291**

risultati campionari

Stima proporzione $p = \frac{\sum c_h}{n} = \frac{210}{300} = 70\%$

$\hat{\text{var}}(\hat{p}_p) = \frac{(1-f)}{n^2} \sum_{h=1}^{H} \frac{n_h^2}{n_h - 1} \hat{p}_h(1-\hat{p}_h)$ trascurando *1-f* e dividendo solo per n_h

Stima varianza $\text{var}(p_p) = \sum \frac{n_h p_h (1-p_h)}{n^2} = 0,0006033$

Stima varianza CCS $\text{var}(p_{cs}) = 0,0007$

$$D\hat{e}ff = \frac{0,0006033}{0,0007} = 0.862$$

$var(p_p)$ 14% più piccola di $var\ (p_{cs})$

Un CCS con $n = \frac{300}{0,862} = 348$ darebbe la stessa varianza di uno stratificato proporzionale con n=300

Stratificato proporzionale e *deff* – esempio 2

Stima spesa annua per abbigliamento delle famiglie italiane nel XXXX
n = 10.000

	CAPOLUOGHI PROVINCIA	ALTRI COMUNI >20 000 AB	COMUNI < 20 000
w_h	0.2	0.3	0.5
$\bar{y}_h$	500	300	220
s_h^2	2500	1600	400

$$\bar{y} = \sum_h w_h \bar{y}_h = 500(0.2) + 300(0.3) + 220(0.5) = 300$$

$$\text{var}(\bar{y}_{prop}) = \frac{1-f}{n} \sum_h w_h s_h^2$$

$$= \frac{1}{10{\cdot}000}[2{\cdot}500(0.2) + 1{\cdot}600(0.3) + 400(0.5)] = \frac{1{\cdot}180}{10{\cdot}000} = 0.118$$

varianza fra gli strati $\frac{1-f}{n} \sum w_h (\bar{y}_h - \bar{y})^2 = \frac{1}{10{\cdot}000}[(500-300)^2 0.2 + (220-300)^2 0.5] = \frac{11{\cdot}200}{10{\cdot}000} = 1.12$

$$\widehat{Deff} = \frac{\text{var}(\bar{y}_{prop})}{\text{var}(\bar{y}_{ccs})} = \frac{0.118}{0.118 + 1.12} = 0.095 \Rightarrow 9.5\%$$

Allocazione di *n* entro gli strati – allocazione NON proporzionale

Varie situazioni in cui allocazione proporzionale non è appropriata:

- vincoli (costi) di indagine
- obiettivi di ricerca
    - strati più variabili
    - elaborazioni per sottopopolazioni (domini di studio)
    - confronti fra strati
- massima precisione delle stime date le risorse (o min costi)

Preferita una allocazione non proporzionale, possibilmente ottima:

$$f_h \propto \frac{s_h}{\sqrt{c_h}} \qquad \text{con} \qquad f_h = \frac{n_h}{N_h}$$

s_h = variabilità strato h

c_h = costo per unità strato h

Allocazione di *n* entro gli strati NON proporzionale

$$n_h = n w_h^*$$

$$w_h^* = \frac{w_h s_h / \sqrt{c_h}}{\sum w_h s_h / \sqrt{c_h}} \quad \text{con} \quad w_h = \frac{N_h}{N} \quad \text{e} \quad \sum_{h=1}^{H} w_h^* = 1$$

più unità negli strati più eterogenei e negli strati meno costosi

se $c_h \simeq c$ per ogni h: allocazione ottima di Neyman

$$n_h = \frac{n w_h s_h}{\sum w_h s_h}$$

più unità negli strati più eterogenei

se $s_h = s \ \forall h \ \Rightarrow \ n_h = \frac{n w_h}{\sum w_h}$ allocazione proporzionale

- conoscenza di s_h
- stimatori pesati con w_h^* (campione non autoponderante)

Stimatore in campione stratificato NON proporzionale

Formule generali in cui si utilizzano gli w_h^* specificati in n_h

$$\bar{y}_o = \sum_h w_h \bar{y}_h$$

varianza campione stratificato

$$var(\bar{y}_{str}) = \sum_h w_h^2 (1 - f_h) \frac{S_h^2}{n_h}$$

se allocazione di Neyman

$$var(\bar{y}_o) = \sum_h w_h^2 \left(n_h^{-1} - N_h^{-1}\right) S_h^2 = \sum_h w_h^2 \frac{\sum_h w_h S_h}{n w_h S_h} S_h^2 - \frac{1}{N} \sum_h w_h S_h^2$$

$$= \frac{1}{n} \left(\sum_h w_h S_h\right)^2 - \frac{1}{N} \sum_h w_h S_h^2$$

si dimostra che (con $1\text{-}f_h \simeq 1\text{-}f$):

$$\hat{\text{var}}(\bar{y}_p) = \underbrace{\hat{\text{var}}(\bar{y}_o)}_{\text{varianza media entro gli strati}} + \underbrace{\frac{1-f}{n} \sum w_h (s_h - \bar{s})^2}_{\text{varianza delle varianze degli strati}} \qquad \bar{s} = \sum w_h s_h$$

$$\hat{\text{var}}(\bar{y}_p) > \hat{\text{var}}(\bar{y}_o)$$

Formazione degli strati - scelta variabili di stratificazione

Non ci sono "criteri oggettivi": dipende da (e quante) informazioni sono disponibili, tenendo conto di alcune condizioni generali e obiettivi stratificazione:

1. Proporzioni $w_h = \frac{N_h}{N}$ note
2. possibilità di selezionare un campione da ogni strato

} non stringenti: post-stratificazione

3. numerosità N_h tale che sia possibile:
   - almeno una selezione per la stima di $\bar{y}_h$
   - almeno due selezioni per la stima di $var(\bar{y}_h)$
4. strati omogenei al loro interno rispetto alle variabili di studio

più variabili *X*, possibilmente non correlate tra loro, combinate anche in modo diverso per definire i vari strati

Variabili di stratificazione: candidate tipiche
attenzione a numero di strati

- ▷ age
- ▷ sex
- ▷ geographical information
- ▷ size of units
- ▷ socio-economic status
- ▷ educational level
- ▷ occupational status
- ▷ type of activity/occupation

- The number of stratifying variables and the number of categories per stratifying variable should not be too large.
  - ▷ age (5)
  - ▷ sex (2)
  - ▷ geographical information (12)
  - ▷ size of units (5)
  - ▷ socio-economic status (4)
  - ▷ educational level (4)
  - ▷ occupational status (4)
  - ▷ type of activity/occupation (5)

- Then, the number of strata is

$$H = 5 \times 2 \times 12 \times 5 \times 4 \times 4 \times 4 \times 5 = 192,000$$

n = 10.000

$$n_h = \frac{10,000}{192,000} = 0.0521$$

Anche utilizzo tecniche di analisi multivariata (clustering, alberi di regressione,…) per individuare il miglior set di variabili di stratificazione
[PISA Italia: *tipo di scuola (5) e area geografica (5)*]

Numerosità campionaria campione stratificato

$$n^* = \frac{s_y^{\,2} Deff(st)}{\mathrm{var}(\bar{y}_{st})} \longleftarrow \mathrm{var}(\bar{y}_{st}) = Deff\,\mathrm{var}(\bar{y}_{cs}) \longrightarrow \frac{s^2}{n}$$

- *Deff* trasferibile fra indagini svolte sulla stessa popolazione
- *Deff* ipotizzabile per proporzioni
- Per indagini multiscopo:
  - selezionare variabili più importanti
  - calcolare allocazione ottima per ogni variabile scelta
  - strato per strato, trovare il compromesso più ragionevole tra le numerosità calcolate (es media o mediana)

Post stratificazione – stratificazione *dopo la* selezione

1. w_h noti $w_h = \frac{N_h}{N}$
2. assegnazione univoca delle unità negli h strati

in assenza della condizione 2) non è possibile mettere in atto la stratificazione,

è possibile, però, stratificare dopo la selezione del campione:

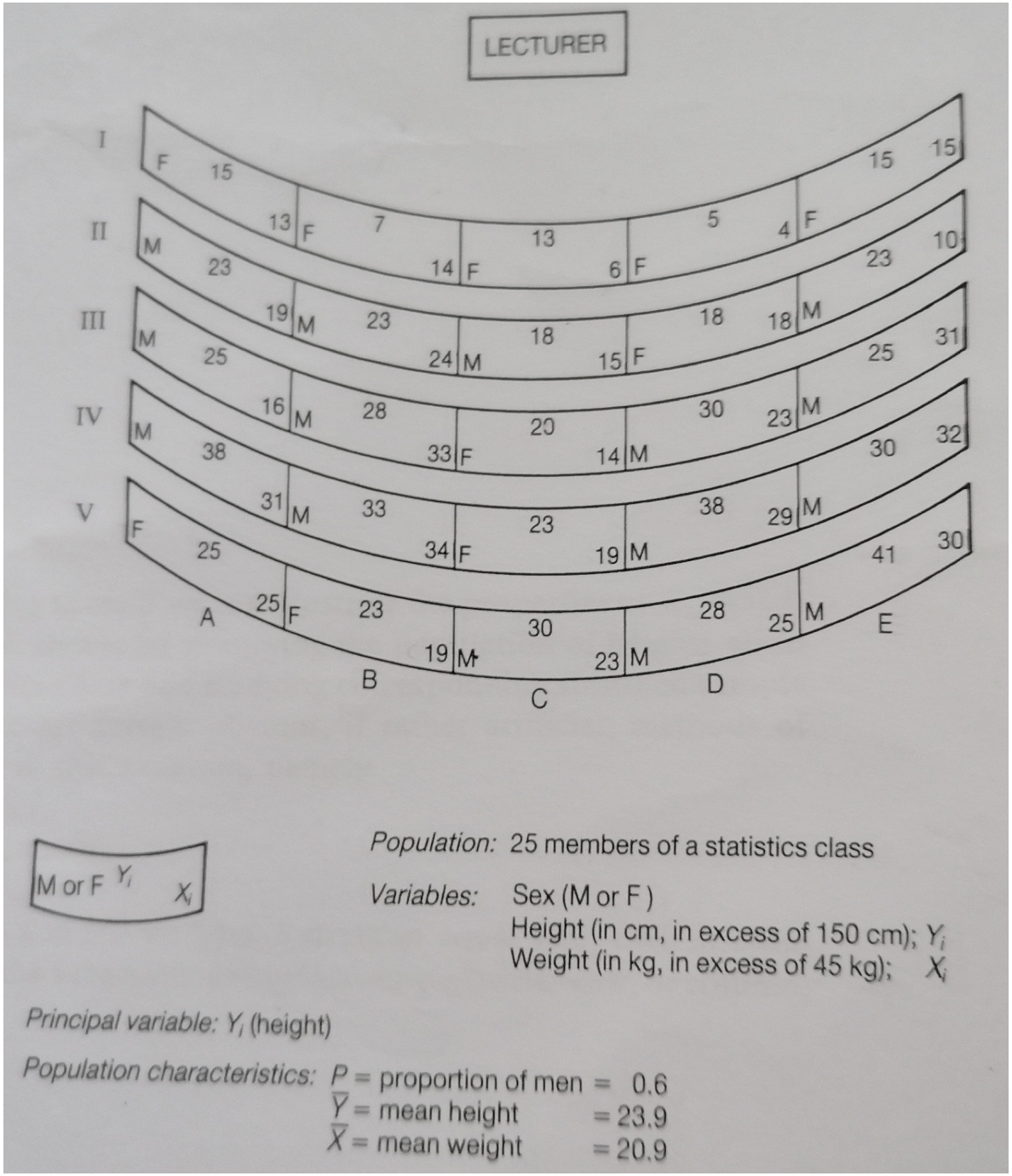
- si seleziona un CCS di n elementi rilevando anche le variabili di stratificazione
- si classifica il campione selezionato in H strati, sulla base delle variabili di stratificazione rilevate
- si usa il peso w_h di ogni strato nella popolazione nella stima

media campionaria

(è uno stimatore rapporto)

$$\bar{y}_{ps} = \sum_h w_h \bar{y}_h = \sum_h w_h \sum_j \frac{y_{hj}}{n_h} \qquad \text{con} \quad w_h = \frac{N_h}{N}$$

Esempio selezione campione (statistics class)



campione n =5 per stimare $\bar{y}$ altezza media

campione accessibile: *studenti 1^ fila* $\bar{y}$ = 55/5 =11

campione ragionato: *studenti in diagonale (IA – VE)* $\bar{y}$ = 27.4

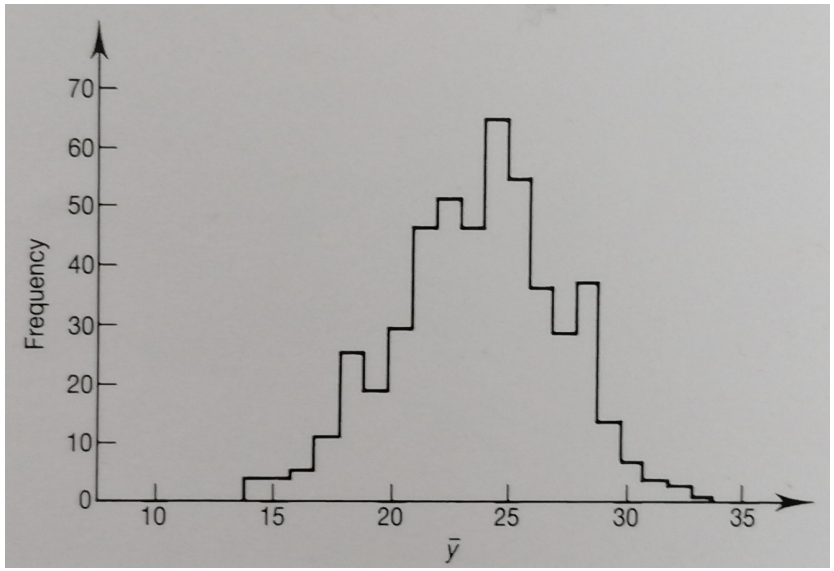
campione CCS: $\bar{y}$ = 27.4 (*var*= 12.9) (y_i= 28, 41, 30, 23, 15)

Campione stratificato (proporzionale):
- 1 studente per fila (riga)
- 1 studente per colonna
- 3 studenti M e 2 F

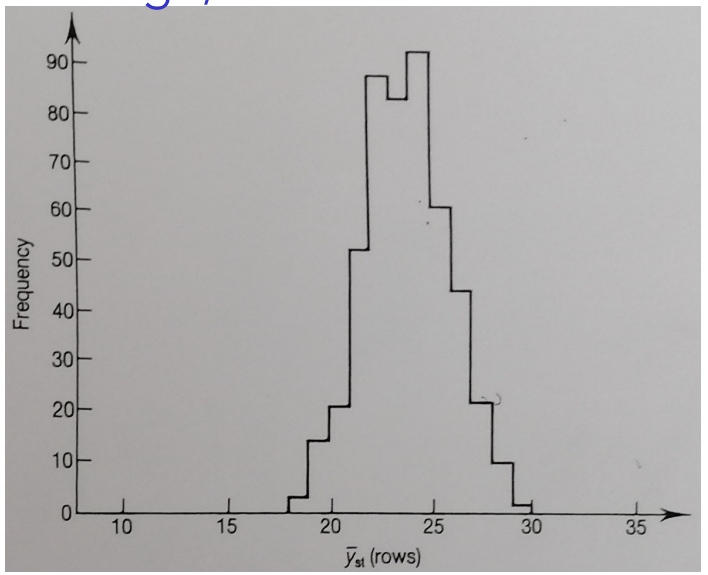
(i)	$\bar{Y}_I = 11.0$	$\bar{Y}_{II} = 21.0$	$\bar{Y}_{III} = 25.6$
	$\bar{Y}_{IV} = 32.4$	$\bar{Y}_V = 29.4$	
	$S_I^2 = 22.0$	$S_{II}^2 = 7.5$	$S_{III}^3 = 14.3$
	$S_{IV}^2 = 39.3$	$S_V^2 = 49.3$	
(ii)	$\bar{Y}_A = 25.2$	$\bar{Y}_B = 22.8$	$\bar{Y}_C = 20.8$
	$\bar{Y}_D = 23.8$	$\bar{Y}_E = 26.8$	
	$S_M^2 = 68.2$	$S_B^2 = 95.2$	$S_C^2 = 39.7$
	$S_D^2 = 161.2$	$S_E^2 = 92.2$	
(iii)	$\bar{Y}_M = 28.9$	$\bar{Y}_F = 16.4$	
	$S_M^2 = 42.0$	$S_F^2 = 45.6$	

Confronti distribuzioni media campionaria (500 replicazioni)

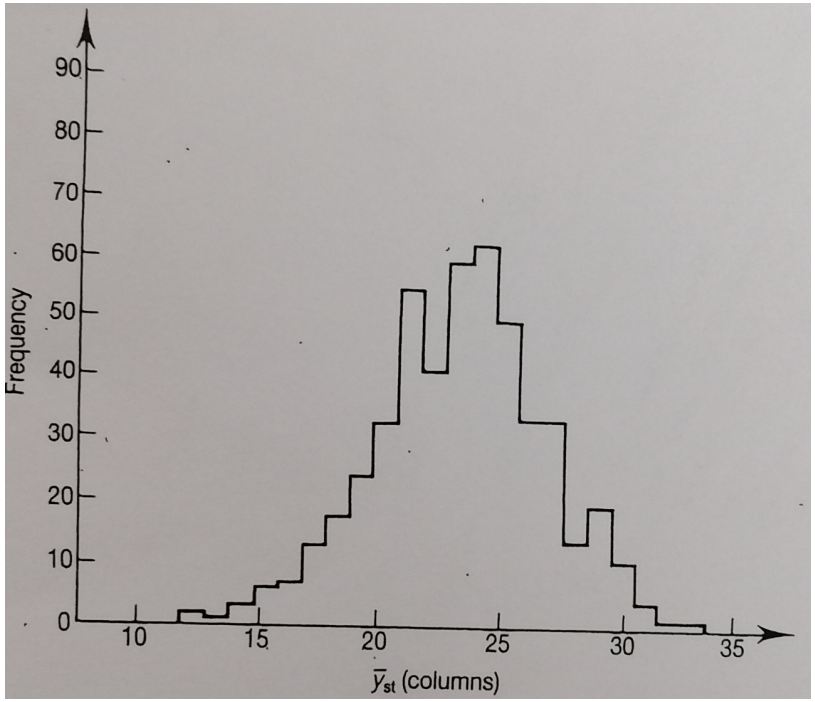
CCS, var = 12.9



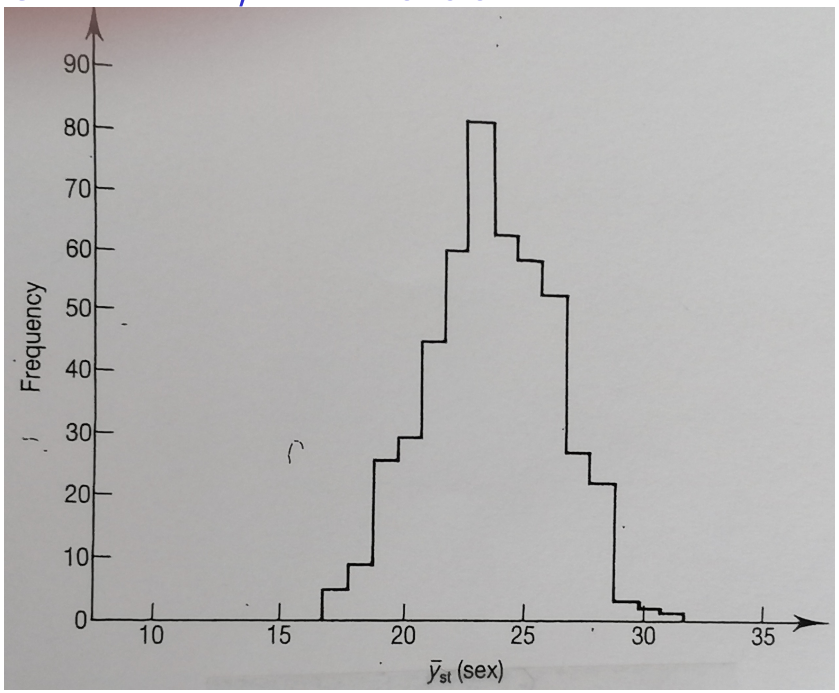
STR riga, var = 4.36



STR colonna, var = 12.99



STR sesso, var = 6.86



Campionamento NON probabilistico

Campionamento non probabilistico:

- non vi è alcun riferimento agli aspetti probabilistici della procedura di selezione e quindi
  - non consente una stima dell'errore di campionamento
  - può introdurre distorsione da selezione dovuta alla non casualità del disegno di campionamento
  - non consente (giustifica) inferenza: i risultati valgono solo per quel campione

Da notare che:

- da un lato, la facilità di realizzazione delle indagini online (apparente rapidità, facilità e costi molto bassi)
- dall'altro lato, la mancanza di appropriate liste della 'popolazione' (sampling frames)

→ fanno sì che molte indagini online sono indagini auto-selezionate.

Ciò può essere vista come una forma di campionamento non probabilistico

Campionamento NON probabilistico – per quote

Campionamento per quote:

- strategia "intermedia" con analogie al campionamento stratificato proporzionale
- la lista di unità della popolazione non è strettamente richiesta

Come si fa un campionamento per quote:

1. Popolazione formata da gruppi definiti in base a variabili di stratificazione (se individui/elettori: sesso, classi di età, titolo di studio, area di residenza, dimensione del comune di residenza) associate a variabile *Y* di interesse.
2. E' necessario conoscere la distribuzione della popolazione (N_h) entro gli strati individuati (da Istat, dati Censimento o stime da indagini campionarie) per calcolare le quote (proporzioni) di popolazione che appartengono ai diversi strati

Campionamento per quote – come si fa

3. Le quote sono usate per distribuire l'ampiezza totale *n* del campione (fissata prima, vedere ampiezza CCS) entro gli strati, ovvero la quota rappresenta il numero di unità da osservare in un determinato strato.
4. Modalità di contatto:

   - libera: ad ogni intervistatore/trice sono assegnate quote di soggetti da intervistare, lasciandoli liberi di contattare chi credono, nel rispetto di tali quote

   - selezione casuale (in indagini CATI) da una lista (elenchi telefonici, panel online/web panel) finché non si raggiunge la quota

   esempio: a ciascun intervistatore (anche CATI) si assegnano 5 soggetti di sesso femminile, tra i 35 e 50 anni, laureati e residenti in comuni di 10.000-50.000 abitanti; 7 soggetti di sesso maschile con le stesse caratteristiche, ecc.