



Tecniche di indagine statistica

Lezione 19



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
 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
- } non stringenti:
post-stratificazione

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

$$n = 10.000$$

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

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

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

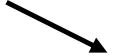
$$n^* = \frac{s_y^2 Deff(st)}{\text{var}(\bar{y}_{st})} \quad \longleftarrow \quad \text{var}(\bar{y}_{st}) = Deff(\text{var}(\bar{y}_{cs})) \quad \longrightarrow \quad \frac{s^2}{n}$$

- *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,

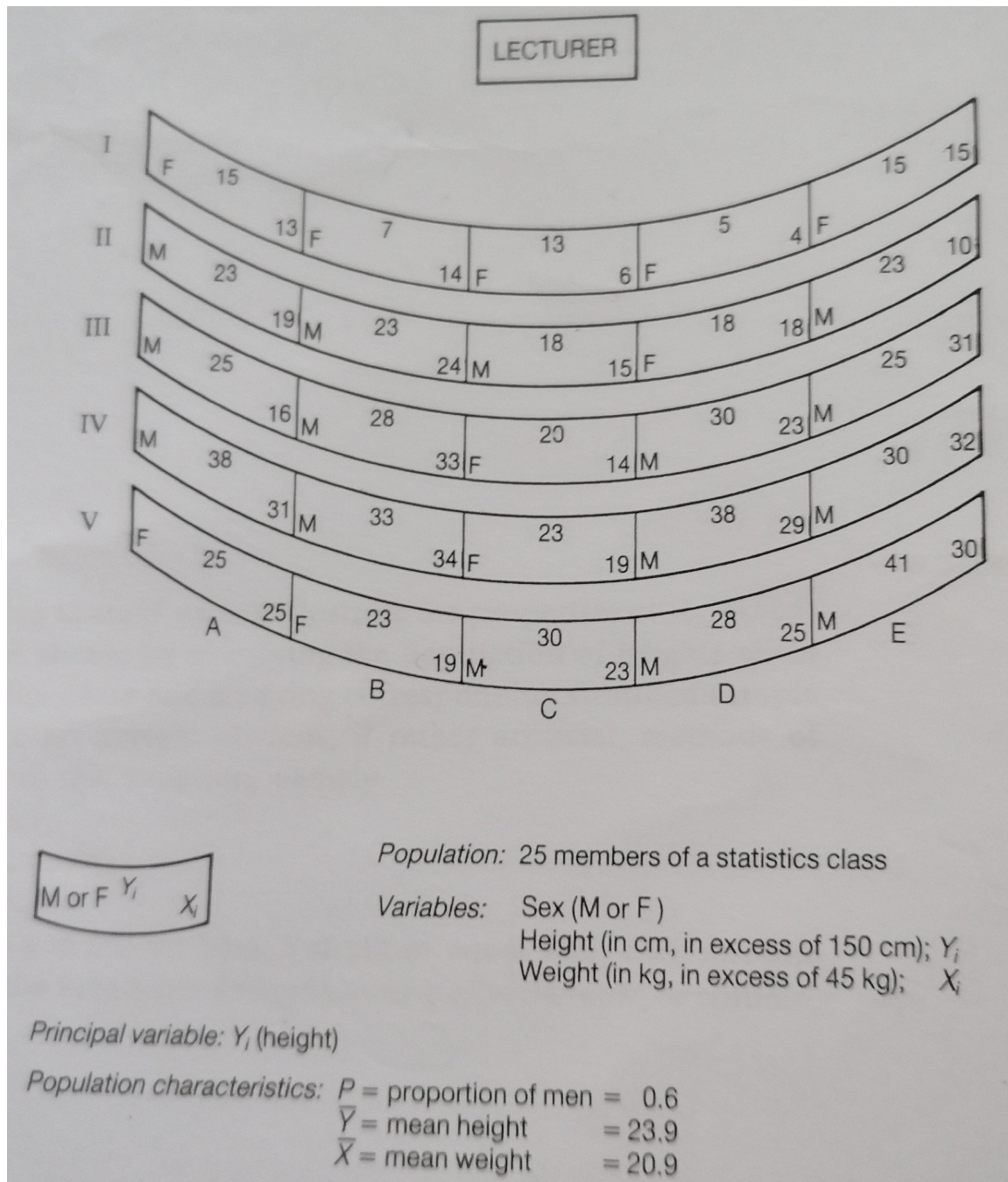
è possibile, però, **stratificare dopo la selezione del campione**:

- 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)

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

Esempio selezione campione (statistics class)



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

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

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

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

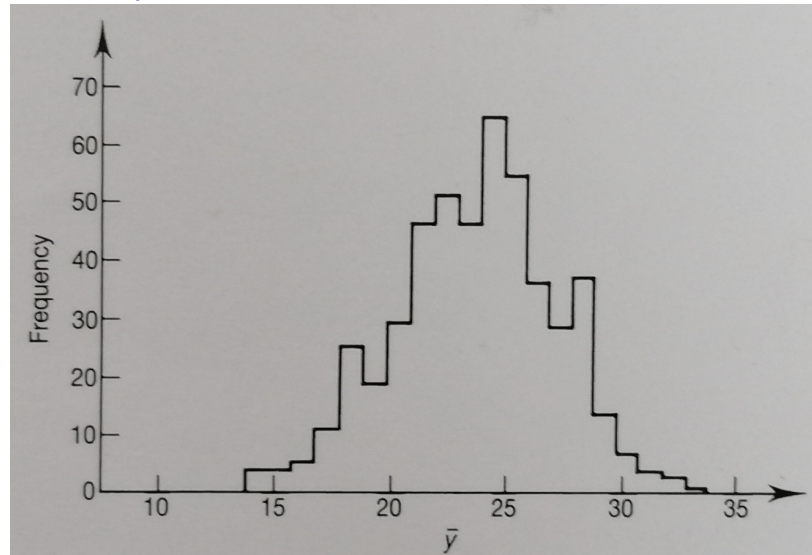
Campione **stratificato** (proporzionale):

- 1 studente per fila (riga)
- 1 studente per colonna
- 3 studenti M e 2 F

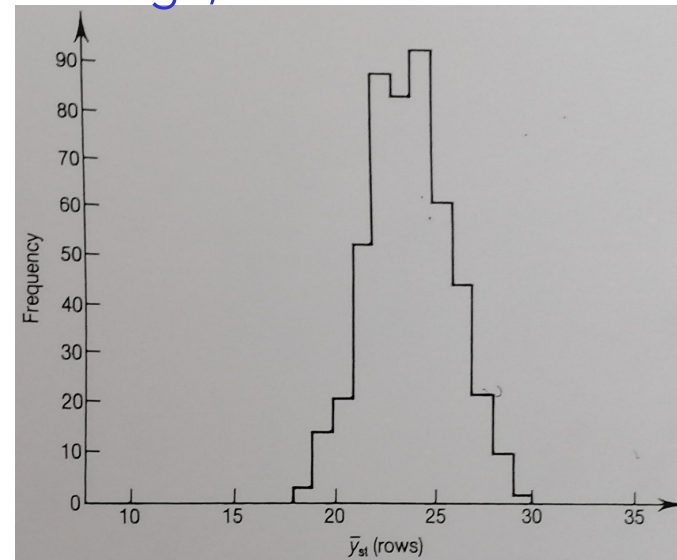
(i)	$\bar{Y}_I = 11.0$ $\bar{Y}_{IV} = 32.4$ $S_I^2 = 22.0$ $S_{IV}^2 = 39.3$	$\bar{Y}_{II} = 21.0$ $\bar{Y}_V = 29.4$ $S_{II}^2 = 7.5$ $S_V^2 = 49.3$	$\bar{Y}_{III} = 25.6$ $S_{III}^3 = 14.3$
(ii)	$\bar{Y}_A = 25.2$ $\bar{Y}_D = 23.8$ $S_M^2 = 68.2$ $S_D^2 = 161.2$	$\bar{Y}_B = 22.8$ $\bar{Y}_E = 26.8$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

Confronti distribuzioni media campionaria (500 repliche)

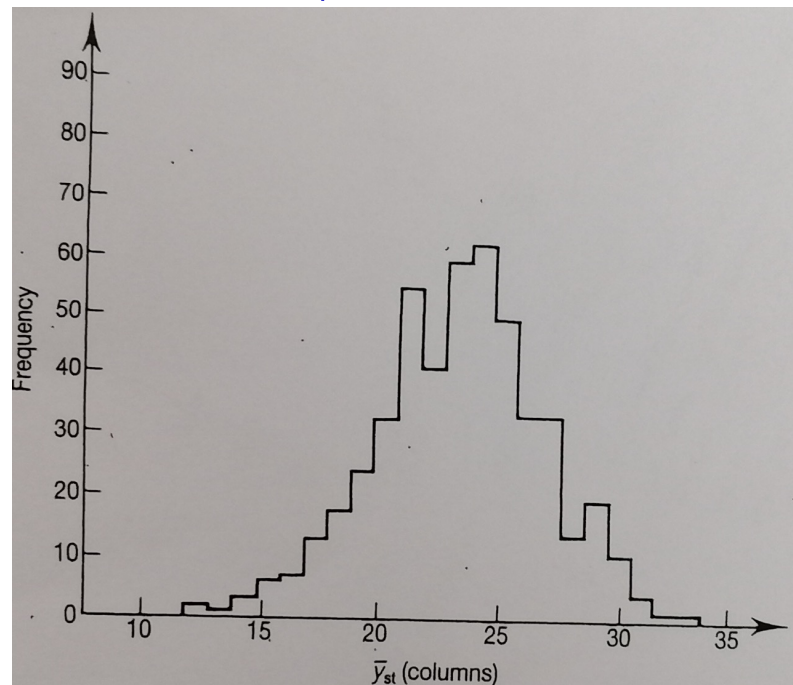
CCS, var = 12.9



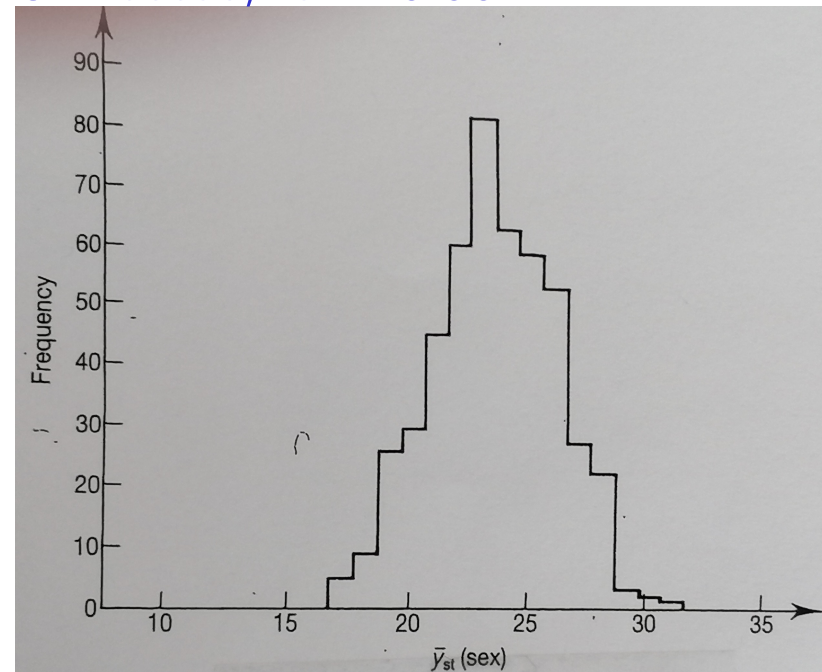
STR riga, var = 4.36



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.