



Tecniche di indagine statistica

Lezione 17 e 18



Pesi (*weights*) di campionamento

π_k = probabilità di inclusione *unità* k nel campione

Possono essere usate per calcolare stime puntuali dei parametri di interesse

ω_k = *peso di campionamento* = $1/\pi_k$ = **Numero** di unità della popolazione *rappresentate dall'unità* k

CCS: $\pi_k = n/N$ da cui $\omega_k = 1/\pi_k = N/n$

CCS pesi tutti uguali: ogni unità nel campione *rappresenta se stessa e altre $N/(n-1)$ unità (non selezionate) della popolazione (in totale N/n)*

$$\sum \omega_k = \sum N/n = N \qquad \sum \omega_k y_k = \sum (N/n) y_k = \hat{t}$$

$$\sum \omega_k y_k / \sum \omega_k = \hat{t} / N = \hat{y}$$

Quando usare un CCS ?

- La popolazione è **omogenea**.
- Si dispone di **buone** liste dell'intera popolazione.
- Il **costo** per raggiungere ogni unità è **omogeneo** e non varia se si prevedesse l'uso di disegni più complessi.
- Si vogliono usare stimatori **semplici**.
- Si vogliono stimare **relazioni complesse** ed altri disegni hanno costi comparabili.

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 :

$$\boxed{\hat{t}_{Y_R} = \hat{B} t_X} \quad \hat{B} = \hat{t}_Y / \hat{t}_X \quad \boxed{\hat{t}_{Y_R} = \hat{t}_Y t_X / \hat{t}_X}$$

in genere non corretti ma più precisi)

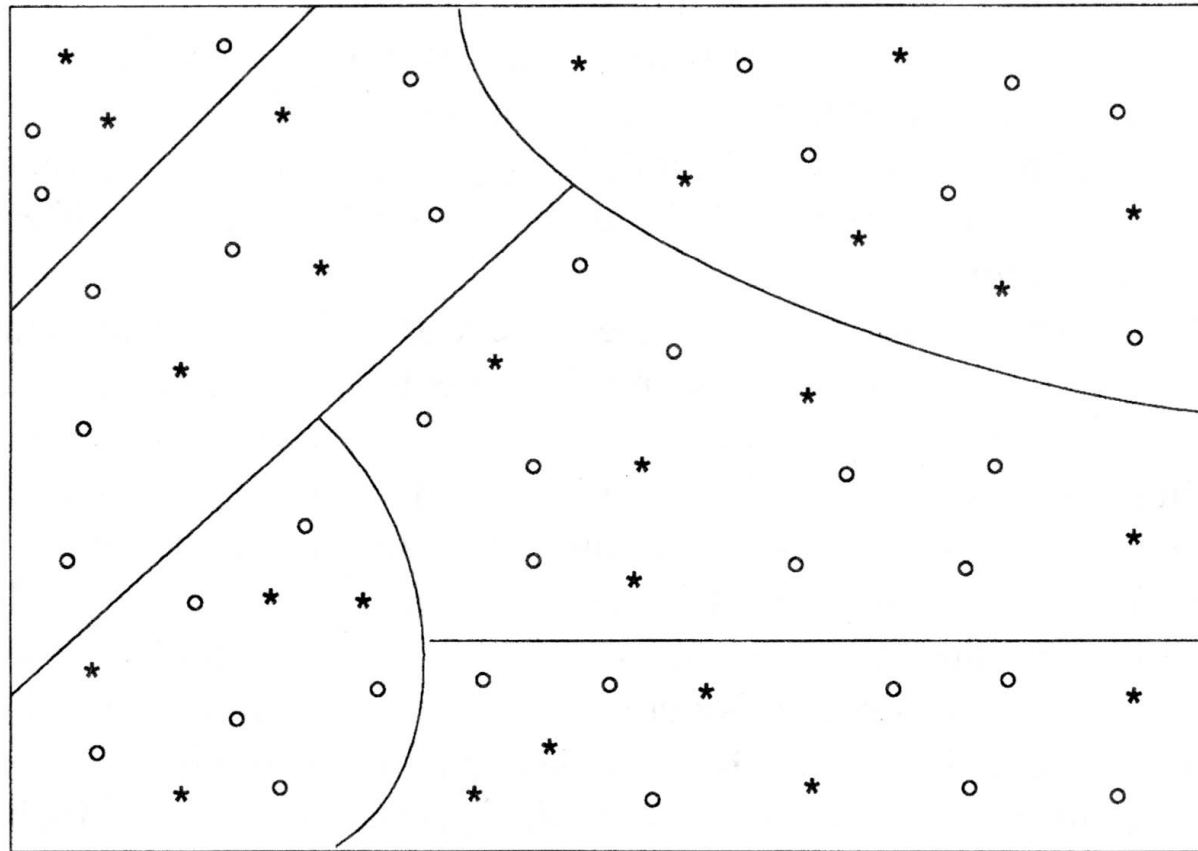
N.b.: campione ancora CCS!

-
- 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 ($\circ$), 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
<u>età</u>	30 anni	30 anni	30 anni	40 anni	40 anni	40 anni	50 anni	50 anni	50 anni
<u>reddito</u>	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 $n_h = 1$ ($h = 1, \dots, 3$) 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:

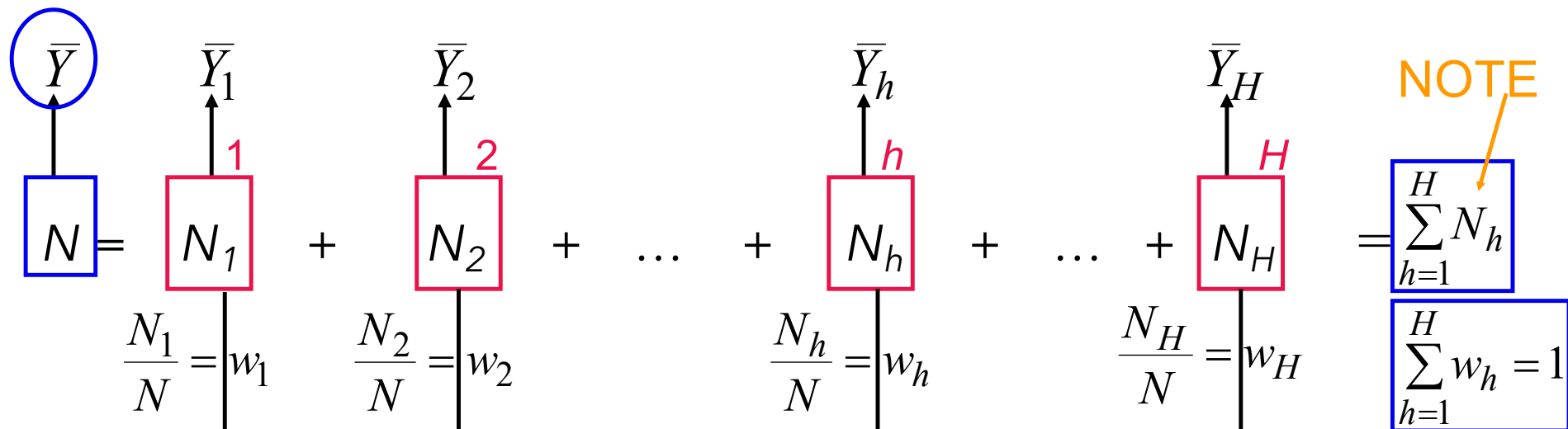
1. “garanzia” contro campioni che, per puro effetto del caso, potrebbero sembrare “poco” rappresentativi
2. stime per gruppi separati (con precisione comparabile)
3. liste disponibili per aree/gruppi separati
4. uso di differenti modalità (costruzione lista, rilevazione, campionamento) nei vari strati 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

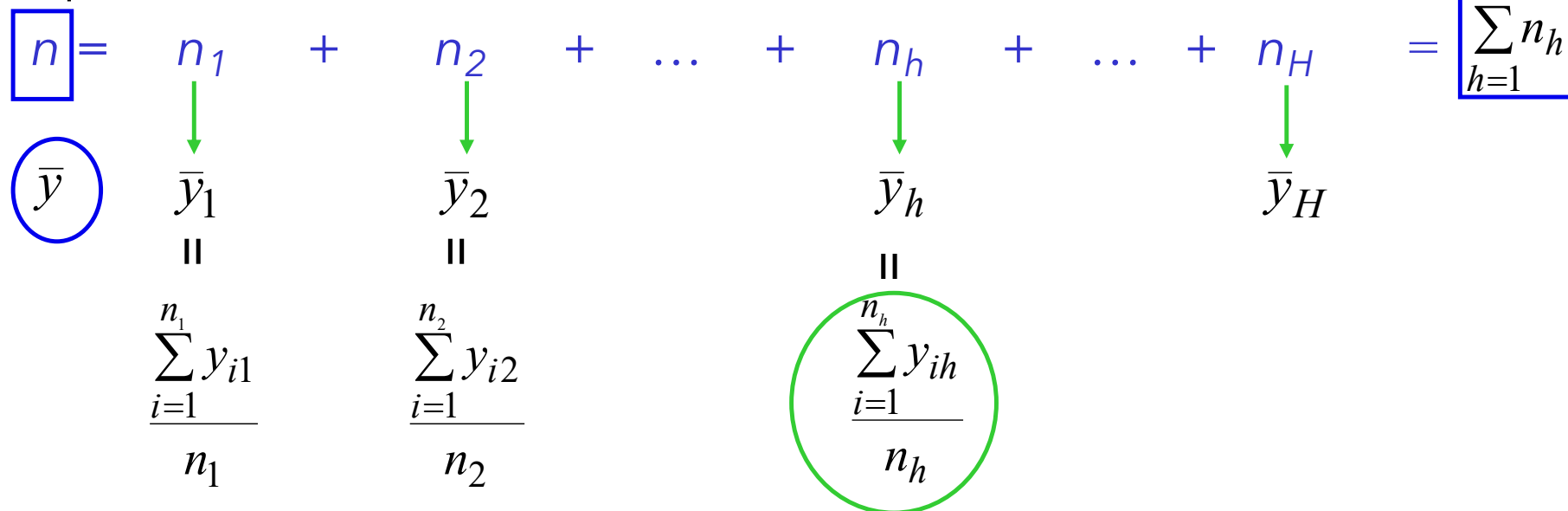
1. La popolazione di N unità è classificata in H strati a seconda di informazioni supplementari
(es. sesso, età, corso di studio, assets bancari, ...)
Classificazione negli strati deve essere **nota prima** della selezione (*ogni unità appartiene a un solo strato*)
2. E' selezionato un campione di numerosità $n_h, h = 1, \dots, H$ 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:



Campione Stratificato – stimatore media pop.ne

Media

Popolazione

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

Media
campionaria

$$\begin{aligned} \bar{y}_{str} &= \sum_{h=1}^H w_h \bar{y}_h \\ &= \sum_{h=1}^H \frac{N_h}{N} \left(\frac{\sum_{i=1}^{n_h} y_{ih}}{n_h} \right) \end{aligned}$$

Varianza
campionaria

$$\widehat{\text{var}}(\bar{y}_{str}) = \sum_{h=1}^H w_h^2 \underbrace{\widehat{\text{var}}(\bar{y}_h)}_{\parallel (1-f_h) \frac{s_h^2}{n_h}} \quad \text{se CCS entro strati}$$

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

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

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

Totale
campionario

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

Varianza
campionaria

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

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

Campione Stratificato – stimatore proporzione

Proporzione campionaria

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

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

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

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)

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

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

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

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

Varianza **media** di strato

Varianza stimatore media campionaria e totale in campione stratificato proporzionale

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

Media
campionaria

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

Totale

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

Stratificazione e efficienza delle stime

Design effect

$$Deff = \frac{\text{var}(z)}{\text{var}(z_{cs})} \begin{matrix} \geq 1 \\ < 1 \end{matrix} \quad \begin{matrix} z \text{ stimatore disegno complesso} \\ z_{cs} \text{ stimatore campione casuale semplice} \end{matrix}$$

con campione stratificato, in generale:

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

poiché:

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

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

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

risultati campionari

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

Stima varianza $\text{var}(p_p) = \sum \frac{n_h p_h (1-p_h)}{n^2} = 0,0006033$ Stima varianza CCS $\text{var}(p_{cs}) = 0,0007$

$$D_{\text{eff}} = \frac{0,0006033}{0,0007} = 0.862 \quad \widehat{\text{var}}(\hat{p}_p) \text{ 14\% più piccola di } \widehat{\text{var}}(\hat{p}_{cs})$$

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 AB
w_h	0.2	0.3	0.5
$\bar{y}_h$	500	300	220
s_h^2	2500	1600	400

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

$$\begin{aligned}\text{var}(\bar{y}_{prop}) &= \frac{1-f}{n} \sum_h w_h s_h^2 \\ &= \frac{1}{10\,000} [2\,500(0.2) + 1\,600(0.3) + 400(0.5)] = \frac{1\,180}{10\,000} = 0.118\end{aligned}$$

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

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

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**:

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

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 $s_h = s \quad \forall h \Rightarrow n_h = \frac{n w_h}{\sum w_h}$ 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

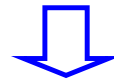
$$var(\bar{y}_{str}) = \sum_h w_h^2 (1 - f_h) \frac{S_h^2}{n_h}$$

se allocazione di Neyman $n_h = \frac{n w_h S_h}{\sum w_h S_h}$

$$\begin{aligned} var(\bar{y}_o) &= \sum_h w_h^2 (n_h^{-1} - N_h^{-1}) S_h^2 = \sum_h w_h^2 \frac{\sum_h w_h S_h}{n w_h S_h} S_h^2 - \frac{1}{N} \sum_h w_h S_h^2 \\ &= \frac{1}{n} (\sum_h w_h S_h)^2 - \frac{1}{N} \sum_h w_h S_h^2 \end{aligned}$$

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

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



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