# Il campionamento casuale stratificato

(ccstr)

# CC Stratificato

- La stratificazione va annoverata tra le tecniche con cui, muovendo da talune conoscenze a priori sulla popolazione, si tende a migliorare l'efficienza del piano di campionamento
  - consente di diminuire l'ordine dei grandezza dell'errore di campionamento senza per questo aumentare la dimensione campionaria
    - si tratta di una sorta di regolamentazione del processo casuale di estrazione delle unità campionarie
- L'ausilio di informazioni supplementari circa uno o più caratteri della popolazione permette, mediante scelta ragionata, di suddividere la popolazione in un certo numero di *strati* il più possibile omogenei nel senso che ciascuno strato presenta variabilità più bassa del carattere considerato.

# CC Stratificato

## Cosa faremo

- Stima della media $\bar{y}_{ST}$ e $V(\bar{y}_{ST})$
- Stima del totale
- Stima di una proporzione

- Allocazione uguale
- Allocazione proporzionale
- Allocazione ottima

*Confronto fra campione casuale stratificato e campione casuale semplice*

# CC Stratificato

Perché la stratificazione? Alcuni vantaggi …

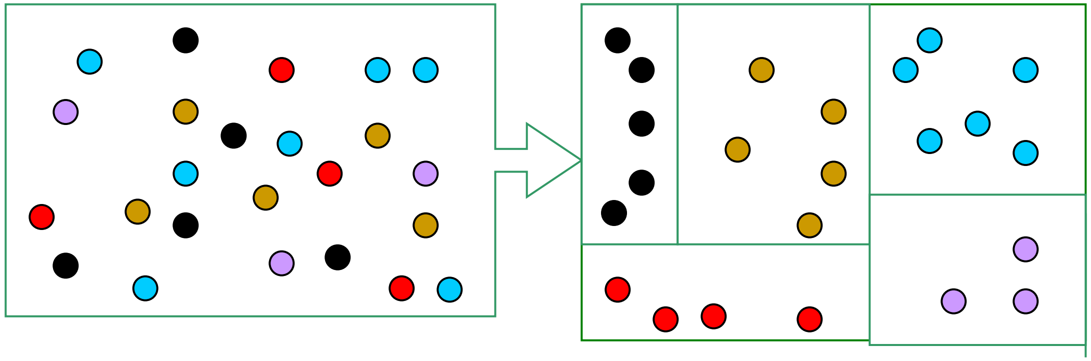
1. **Per ridurre la varianza della stima (miglioramento della precisione)**
   - Se vi sono differenze notevoli fra i vari strati

2. **Ogni strato può essere formato in modo che all'interno si usano metodi e procedure diverse**
   - Tutti gli strati rappresentati
     - **Statistiche differiscono in modo sostanziale**
       - **Militari**
       - **Gruppi etnici**
     - **Differenze nelle liste per differenti tipologie**
       - **Proprietari telefono**
       - **Affittuari**
   - **La diversa tipologia degli elementi della popolazione può richiedere procedure differenti**
     - Operai
     - Impiegati

3. **Gli strati sono tali che le sottopopolazioni corrispondenti sono e costituiscono i domini di studio**

# CC Stratificato

Perché la stratificazione? Alcuni vantaggi …



# CC Stratificato

- **Definizione:**
  - Si suddivida la popolazione iniziale formata da N unità in L sottopopolazioni o strati, all'interno dei quali le unità siano omogenee secondo qualche criterio.

$$N = N_1+N_2+N_3+\ldots+N_L$$

  - Da ogni strato si estragga un campione casuale semplice

*Tale campionamento dà luogo al piano campionario Casuale Stratificato*

# CC Stratificato

Per ottenere i vantaggi del ccsstr devono:

- essere note le dimensioni di ogni strato $N_h$
- essere indipendenti gli strati $N_h$
- essere omogenee le sottopopolazioni individuate $N_h$ cioè
  - Unità omogenee rispetto alla variabile oggetto d'indagine

# CC Stratificato

**Esempio di utilità della stratificazione:**

27 maschi 13 femmine N = 40

Estraggo un campione casuale semplice n = 4

$$\binom{27}{4} = 17550 \quad \text{tutti da maschi}$$

$$\binom{N}{n} = \binom{40}{4} = \frac{40 \times 39 \times 38 \times 37}{4 \times 3 \times 2} = 91390$$

$$\binom{13}{4} = 715 \quad \text{tutti da femmine}$$

Se ho un campione formato solo da maschi avrò una sovrastima della statura media

Se stratifico per il genere, avrò una stima più equilibrata. Non ho campioni di sole femmine o di soli maschi.

# CC Stratificato

Unità omogenee rispetto alla variabile oggetto d’indagine

Omogeneità

In senso concettuale
Esempio: secondo le dimensioni delle città

In senso tecnico
Le unità in ogni strato presentano varianza (rispetto alla caratteristica studiata) più piccola

Le due omogeneità dovrebbero coincidere

# CC Stratificato

Omogeneità in senso tecnico
Le unità in ogni strato presentano varianza (rispetto alla caratteristica studiata) più piccola

Se invece di considerare la popolazione in un’unica maniera, la si suddivide in più parti, *sottopopolazioni*, il più possibile omogenee tra loro (con piccola variabilità)

$$\begin{tikzcd}
 & S^2 \arrow[dl] \arrow[d] \arrow[dr] & \\
S_1^2 & S_2^2 \quad \ldots & S_L^2
\end{tikzcd}$$

A parità di condizioni, se gli strati hanno variabilità ridotta ( $S_1^2 \ldots S_L^2$ il più piccolo possibile), si intuisce che la stima che se ne ricava da ogni strato sarà molto più precisa e, quindi, anche quella complessiva sarà molto più precisa.

# CC Stratificato

**Come si formano gli strati?**

*Deve essere resa massima la differenza fra le medie degli strati, cioè massima l'eterogeneità fra gli strati.*

Se si ritiene opportuno, la stratificazione può essere fatta sulla base di più caratteri contemporaneamente $k_1$ e $k_2$. Se sono i livelli o le modalità dei due caratteri sulla cui base viene effettuata la stratificazione, il numero di strati risultanti sarà uguale al prodotto $k_1 \times k_2$

# CC Stratificato

**Notazioni**

L= numero degli strati

h= indicatore generico strato

$N_h$ = Numerosità all'interno degli strati della popolazione

$n_h$ = Numerosità all'interno degli strati del campione

$$W_h = \frac{N_h}{N}$$

Peso dello strato nella popolazione

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

Peso dello strato nel campione

$$f_h = \frac{n_h}{N_h}$$

Frazione di campionamento nello strato

$$\sum_{h=1}^{L} N_h = N \qquad \sum_{h=1}^{L} n_h = n$$

# CC Stratificato

**Notazioni**

$Y_{hi}$ Valore della variabile oggetto di studio per l'unità i nello strato h nella popolazione

$y_{hi}$ Valore della variabile oggetto di studio per l'unità i nello strato h nel campione

$$\overline{Y}_h = \frac{\sum_{i=1}^{N_h} Y_{hi}}{N_h}$$

Media della popolazione nello strato h

$$\overline{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$$

Media del campione nello strato h

# CC Stratificato

**Notazioni**

$$S_h^2 = \frac{\sum_{i=1}^{N_h} \left(Y_{hi} - \overline{Y}_{hi}\right)^2}{N_h - 1}$$

Varianza nello strato h della popolazione

$$s_h^2 = \frac{\sum_{i=1}^{n_h} \left(y_{hi} - \overline{y}_{hi}\right)^2}{n_h - 1}$$

Varianza nello strato h del campione

# CC Stratificato

**Spazio campionario**

Numero dei possibili campioni senza reimmissione

$$\binom{N_1}{n_1} \times \binom{N_2}{n_2} \times \ldots\ldots \times \binom{N_L}{n_L}$$

Sono campioni indipendenti
e
estrazioni indipendenti

Numero dei possibili campioni con reimmissione

$$N_1^{\,n_1} \times N_2^{\,n_2} \times \ldots\ldots \times N_h^{\,n_h}$$

Sono campioni indipendenti
e
estrazioni indipendenti

# CC Stratificato

**Allocazione delle unità**

In genere il criterio all'interno del quale ci si muove per scegliere la dimensione del campione e quindi degli strati è quello di soddisfare una delle seguenti condizioni:

- *rendere minima la varianza* dello stimatore in funzione di un dato costo
- *minimizzare il costo dell'indagine* per un prefissato valore della varianza

**Allocazione uguale**

L'allocazione può essere uguale e ciò comporta che si estragga un numero identico di unità in ogni strato, ovvero:

$$n_1 = n_2 = \ldots\ldots = n_L$$

In questo caso gli strati sono bilanciati, per cui: $n_h = \dfrac{n}{L}$

# CC Stratificato

**Allocazione proporzionale**

Al contrario, l'allocazione può essere proporzionata alla numerosità dello strato nella popolazione, ovvero:

$$n_h \propto W_h$$

per cui se fissiamo la dimensione campionaria n si avrà per ogni strato:

$$n_h = n \times \frac{N_h}{N} \qquad \text{per h=1,...L}$$

# CC Stratificato

**Allocazione ottima**

La dimensione dello strato è funzione di:

$$n_h \rightarrow \left(N_h ; S_h^2\right)$$

strati molto numerosi e omogenei al loro interno rispetto al carattere di interesse, sono rappresentati, nel campione, da poche unità;
strati poco numerosi e fortemente variabili al loro interno necessitano di numerose unità nel campione.

$$n_h = n \times \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}$$

# CC Stratificato

**Allocazione ottima**

È possibile tenere conto anche dei costi di ogni singola unità

La funzione costo più semplice che viene usualmente adottata in letteratura è la seguente:

$$C = C_0 + \sum_{h=1}^{L} C_h n_h$$

$C_0$ denota i costi fissi dell'indagine, quelli cioè che dipendono dalla strategia. Fanno parte di questa categoria quelli conseguenti al reperimento delle informazioni ausiliarie, alla formazione degli strati, alla preparazione delle liste, degli intervistatori, ecc…

$C_h$ invece rappresenta il costo complessivo da sostenere per ogni unità del campione nello strato h. Si chiama costo variabile perché in genere varia di strato in strato.

# CC Stratificato

**Allocazione ottima**

Si possono considerare funzioni di costo più complesse e specifiche che meglio possono identificare i costi effettivi.
Ad esempio, se il costo maggiore è costituito dalle spese di viaggio dei rilevatori, la funzione più appropriata è:

$$C = \sum_{h=1}^{L} t_h \sqrt{n_h}$$

$t_h$ è il costo del viaggio per unità

Se consideriamo la funzione lineare dei costi, le dimensioni degli strati nel campione dipenderanno anche da tali costi nel seguente modo:

$$n_h = n \times \frac{W_h S_h / \sqrt{C_h}}{\sum_{h=1}^{L} W_h S_h / \sqrt{C_h}}$$

se i costi sono bassi si estraggono più unità da quello strato

# CC Stratificato

**Probabilità di inclusione del primo ordine**

Probabilità che l'i-esima unità entri a far parte dello strato h-esimo

$$\pi_{hi} = \frac{n_h}{N_h}$$

se si verifica che:

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \dots\dots = \frac{n_L}{N_L}$$

si ha una *allocazione proporzionale* e il piano è **autoponderante**

# CC Stratificato

**Probabilità di inclusione del secondo ordine**

**rispetto ad h dello stesso strato**
Probabilità che l'i-esima unità e la j-esima unità entrino a far parte dello stesso strato h-esimo

$$\pi_{ij} = \frac{n_h}{N_h}\frac{n_h - 1}{N_h - 1} \qquad \forall \ i \neq j$$

**rispetto a due strati diversi**
Probabilità che l'i-esima unità entri a far parte del campione nello strato h-esimo e la j-esima unità entri a far parte del campione nello strato h'-esimo

$$\pi_{(hi)(h'j)} = \frac{n_h}{N_h}\frac{n_{h'}}{N_{h'}}$$

# CC Stratificato

**Stima dei parametri**

*Media*

Se l'interesse è rivolto alla media della variabile Y in una popolazione generica suddivisa in L strati, cioè:

$$\overline{Y}_{ST} = \frac{\sum_{h=1}^{L} N_h \overline{Y}_h}{N} = \sum_{h=1}^{L} W_h \overline{Y}_h$$

Media della Y nella popolazione ottenuta come media aritmetica ponderata delle medie in ogni strato di numerosità $N_h$

si potrebbe usare come stimatore *la media campionaria*, ottenuta anch'essa come media aritmetica ponderata di tutte le medie in ogni strato del campione:

$$\overline{y}_{ST} = \frac{\sum_{h=1}^{L} N_h \overline{y}_h}{N} = \sum_{h=1}^{L} W_h \overline{y}_h$$

# CC Stratificato

## Stima dei parametri

*Media*

La media della variabile Y nel campione, invece è data dalla:

$$\bar{y} = \frac{\sum_{h=1}^{L} n_h \bar{y}_h}{n} = \sum_{h=1}^{L} w_h \bar{y}_h$$

i pesi sono costruiti tenendo conto delle dimensioni degli strati nel campione. La media campionaria e quella del campione coincidono solo se

$$\frac{n_h}{n} = \frac{N_h}{N} \text{ o } \frac{n_h}{N_h} = \frac{n}{N} \text{ o } f_h = f$$

cioè, solo nel caso di allocazione proporzionale, ovvero di campionamento autoponderante.

# CC Stratificato

## *Media*

**Teorema 1:** Dato un campione stratificato, non necessariamente casuale, se in ogni strato la stima campionaria della media $\bar{y}_h$ è non distorta, allora $\bar{y}_{ST}$ è una stima non distorta di $\bar{Y}_{ST}$

**Dimostrazione:**

$$E(\bar{y}_{ST}) = E\left(\sum_{h=1}^{L} W_h \bar{y}_h\right) = \sum_{h=1}^{L} W_h E(\bar{y}_h) = \sum_{h=1}^{L} W_h \bar{Y}_h = \bar{Y}$$

infatti vale che:

$$\bar{Y} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{N} Y_{hi}}{N} = \frac{\sum_{h=1}^{L} N_h \bar{Y}_h}{N} = \sum_{h=1}^{L} W_h \bar{Y}_h$$

# CC Stratificato

## *La varianza della Media*

**Teorema 2:** Dato un campione stratificato, non necessariamente casuale, se i campioni sono estratti in modo indipendente nei differenti strati, si ha:

$$V(\bar{y}_{ST}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h)$$

Dove $V(\bar{y}_h)$ è la varianza della stima della media per lo strato h. Per calcolare la varianza della media è necessario che in ogni strato ci siano almeno due unità.

**Dimostrazione:**

poiché $\bar{y}_{ST} = \sum_{h=1}^{L} W_h \bar{y}_h$ cioè è una combinazione lineare delle $\bar{y}_h$ con pesi dati da $W_h$

possiamo scrivere la varianza in termini di combinazione lineare come varianza di una somma:

$$V(\bar{y}_{ST}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h) + 2\sum_{h=1}^{L}\sum_{j>h}^{L} W_h W_j \operatorname{cov}(\bar{y}_h \bar{y}_j)$$

# CC Stratificato

## *La varianza della Media*

**Teorema 2:** Dato un campione stratificato, non necessariamente casuale, se i campioni sono estratti in modo indipendente nei differenti strati, si ha:

**Dimostrazione:**

$$V(\bar{y}_{ST}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h) + 2\sum_{h=1}^{L}\sum_{j>h}^{L} W_h W_j \operatorname{cov}(\bar{y}_h \bar{y}_j)$$

Poiché le estrazioni sono indipendenti, le covarianze sono tutte nulle, per cui

$$V(\bar{y}_{ST}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h)$$

# CC Stratificato

## *La varianza della Media*

I risultati dei teoremi 1 e 2 si possono riassumere nel seguente modo:

Se $\bar{y}_h$ è una stima non distorta di $\bar{Y}_h$ in ogni strato e la selezione dei campioni è indipendente per ogni strato, allora la media campionaria è uno stimatore non distorto del parametro della popolazione

$$E\left(\bar{y}_{ST}\right) = \bar{Y}$$

(risultato del teorema 1)

con varianza pari a:

$$V\left(\bar{y}_{ST}\right) = \sum_{h=1}^{L} W_h^2 V\left(\bar{y}_h\right)$$

(risultato del teorema 2)

Ciò significa che la varianza della media campionaria dipende esclusivamente dalla varianza delle stime delle medie dei singoli strati $\bar{y}_h$

se fosse possibile suddividere la popolazione altamente variabile in strati omogenei, con varianza nulla all'interno di ogni strato, potremmo stimare la media della popolazione senza errore

# CC Stratificato

## *La varianza della Media*

**Teorema 3:** Se i campioni sono estratti casualmente, si ha:

$$V(\bar{y}_{ST}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h (N_h - n_h)\frac{S_h^2}{n_h} = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}(1 - f_h)$$

**Dimostrazione:**

Ci si avvale dei risultati del precedente teorema

$$V(\bar{y}_{ST}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 V(\bar{y}_h) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 \frac{S_h^2}{n_h}\left(\frac{N_h - n_h}{N_h}\right) =$$

$$\frac{1}{N^2}\sum_{h=1}^{L} N_h (N_h - n_h)\frac{S_h^2}{n_h} = \frac{1}{N^2}\left[\sum_{h=1}^{L} N_h N_h \frac{S_h^2}{n_h} - \sum_{h=1}^{L} N_h n_h \frac{S_h^2}{n_h}\right] = \frac{1}{N^2}\left[\sum_{h=1}^{L} N_h^2 \frac{S_h^2}{n_h} - \sum_{h=1}^{L} \frac{N_h}{N_h} N_h n_h \frac{S_h^2}{n_h}\right] =$$

$$= \frac{1}{N^2}\left[\sum_{h=1}^{L} N_h^2 \frac{S_h^2}{n_h} - \sum_{h=1}^{L} \frac{N_h^2}{N_h} n_h \frac{S_h^2}{n_h}\right] = \sum_{h=1}^{L} \frac{N_h^2}{N^2}\left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h} = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}(1 - f_h)$$

dove $V(\bar{y}_h) = \frac{S_h^2}{n_h}\left(\frac{N_h - n_h}{N_h}\right) = \frac{S_h^2}{n_h}(1 - f_h)$

# CC Stratificato

## *La varianza della Media*

**Osservazioni**

Se $f_h = \frac{n_h}{N_h} \longrightarrow 0$

possiamo ignorare la correzione per popolazione finita perché

$1 - f_h$ tende a 1, quindi possiamo scrivere

$$V(\bar{y}_{ST}) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h^2 S_h^2}{n_h} = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}$$

# CC Stratificato

## *La varianza della Media*

**Allocazione uguale**

$n_h = \frac{n}{L}$ e la varianza dello stimatore media assume la seguente forma:

$$V\left(\bar{y}_{STug}\right) = \frac{L}{n}\sum_{h=1}^{L} W_h^2 S_h^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

dove L è il numero di strati. Infatti

$$V\left(\bar{y}_{STug}\right) = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h} \frac{N_h - n_h}{N_h}$$

$$V\left(\bar{y}_{STug}\right) = \sum_{h=1}^{L} W_h^2 S_h^2 \frac{L}{n} \frac{N_h - \frac{n}{L}}{N_h} = \frac{L}{n}\sum_{h=1}^{L} W_h^2 S_h^2 \frac{LN_h - n}{LN_h}$$

# CC Stratificato

## *La varianza della Media*

**Allocazione proporzionale**

tenendo presente che $f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$ e che quindi $n_h = n\frac{N_h}{N}$

con le opportune sostituzioni si ottiene:

$$V(\bar{y}_{ST}) = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}(1 - f_h) = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{nN_h} N\left(1 - \frac{n_h}{N_h}\right) =$$

$$= \sum_{h=1}^{L} \frac{N_h^2}{N^2} \frac{S_h^2}{nN_h} N\left(1 - \frac{n}{N}\right) = \sum_{h=1}^{L} \frac{S_h^2}{nN} N_h (1 - f) = \sum_{h=1}^{L} \frac{S_h^2}{N} N_h \left(\frac{1-f}{n}\right) =$$

$$= \sum_{h=1}^{L} W_h S_h^2 \left(\frac{1-f}{n}\right) = \frac{1-f}{n} \sum_{h=1}^{L} W_h S_h^2$$

# CC Stratificato

## *La varianza della Media*

**Allocazione proporzionale**

Se tutte le varianze all'interno degli strati sono uguali $S_h^2 = S_w^2$

allora:

$$V(\bar{y}_{ST}) = \frac{S_w^2}{n} \frac{N-n}{N}$$

infatti, sostituendo si ha:

$$V(\bar{y}_{ST}) = \frac{1-f}{n} \sum_{h=1}^{L} \frac{N_h S_h^2}{N} = \frac{1-\frac{n}{N}}{n} \frac{1}{N} S_w^2 \sum_{h=1}^{L} N_h =$$

$$= \frac{N-n}{nN} \frac{1}{N} N S_w^2 = \frac{N-n}{N} \frac{S_w^2}{n} = \frac{S_w^2}{n}(1-f)$$

coincide con la varianza della media nel caso di campionamento casuale semplice quindi la stratificazione non migliora la precisione delle stime

# CC Stratificato

## *La varianza della Media*

**Allocazione proporzionale**

La maggior parte delle volte, non si conosce la varianza della variabile Y negli strati della popolazione

$S_h^2$

in questi casi occorre stimarla. Se in ogni strato estraiamo un campione casuale, costituito da almeno 2 unità, una stima non distorta di $S_h^2$ è:

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

Quindi $V(\bar{y}_{ST})$ quando $S_h^2$ non è nota si stima con:

$$v(\bar{y}_{ST}) = s^2(\bar{y}_{ST}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h (N_h - n_h) \frac{s_h^2}{n_h}$$

# CC Stratificato

## *La varianza della Media*

**Allocazione proporzionale**

$$v(\bar{y}_{ST}) = s^2(\bar{y}_{ST}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{s_h^2}{n_h}$$

che può essere scritta anche nel seguente modo:

$$s^2(\bar{y}_{ST}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{s_h^2}{n_h} = \sum_{h=1}^{L}\frac{1}{N^2}\left(N_h^2 - N_h n_h\right)\frac{s_h^2}{n_h} =$$

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

dove l'ultimo termine della somma è il fattore di correzione per popolazione finita

# CC Stratificato

## *La varianza della Media*

**Allocazione ottima**

Nel caso di allocazione ottima occorre tenere presente che a parità di n gli $n_h$ devono essere individuati tenendo conto dei **costi** e della **precisione** che si vuole garantire allo stimatore. Nella determinazione degli $n_h$ solo uno dei vincoli può essere rispettato. Indipendentemente dal vincolo considerato, possiamo considerare la formula generale della varianza:

Indipendentemente dal vincolo considerato, possiamo considerare la formula generale della varianza:

$$V(\bar{y}_{ST}) = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}(1 - f_h) = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h S_h^2}{N}$$

che a seconda degli $n_h$ ottenuti minimizzando la varianza o i costi, assumerà la forma:

$$V(\bar{y}_{ST}) = \frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{n} - \frac{\sum_{h=1}^{L} W_h S_h^2}{N}$$

Essa è ottenuta dalla formula della varianza generale nella quale vengono sostituiti gli $n_h$ ottimi

# CC Stratificato

## *La varianza della Media*

**Determinazione degli $n_h$ ottimi e della dimensione campionaria**

Assumendo valida la funzione di costo lineare se la dimensione degli strati è proporzionale alla numerosità dello strato, alla variabilità interna allo strato e inversamente proporzionale ai costi, cioè se

$$n_h \propto \frac{W_h S_h}{\sqrt{c_h}}$$

allora **la varianza della media campionaria è un minimo per un costo prefissato e il costo è minimo per una specificata varianza**.

L'individuazione degli $n_h$ può avvenire o minimizzando i costi per una varianza data o minimizzando la varianza per certi costi.
Ovviamente è possibile, come nel ccs, minimizzare la varianza per un n fissato.

# CC Stratificato

*Individuazione degli $n_h$*
*Minimizzazione della varianza fissando la dimensione campionaria **n***
***Allocazione ottima di Neyman***

Procedendo inizialmente tentando di minimizzare la varianza, ci si trova nelle seguenti condizioni:

$V(\bar{y}_{ST})$ minima con il vincolo che $\sum_{h=1}^{L} n_h = n$

Si tratta di un problema di *minimo vincolato*, per risolvere il quale usiamo i Moltiplicatori di Lagrange:

$$F = F_0 + \lambda_1 F_1 + \lambda_2 F_2 + \lambda_3 F_3 + \ldots \lambda_k F$$

con k vincoli. Dato che nel nostro caso è presente solo un vincolo, sostituendo nella precedente formula

$F_0$ $F_1$ e $\lambda$ con le quantità corrispondenti cioè:

$$F_0 = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2 \qquad \lambda_1 F_1 = \lambda\left(\sum_{h=1}^{L} n_h - n\right)$$

$$F = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2 + \lambda\left(\sum_{h=1}^{L} n_h - n\right)$$

# CC Stratificato

Per trovare il minimo di tale funzione, bisogna calcolare le derivate parziali rispetto a $n_h$ e a $\lambda$ e porle uguali a zero.

$$\frac{\partial F}{\partial n_h} = -\frac{W_h^2 S_h^2}{n_h^2} + \lambda = 0 \Rightarrow \lambda = \frac{W_h^2 S_h^2}{n_h^2} \Rightarrow n_h^2 = \frac{W_h^2 S_h^2}{\lambda} \Rightarrow n_h = \frac{W_h S_h}{\sqrt{\lambda}}$$

e

$$\frac{\partial F}{\partial \lambda} = \sum_{h=1}^{L} (n_h) - n = 0 \Rightarrow n = \sum_{h=1}^{L} n_h = \sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{\lambda}} \Rightarrow \sqrt{\lambda} = \frac{1}{n} \sum_{h=1}^{L} W_h S_h$$

Da cui, sostituendo nella prima equazione $\sqrt{\lambda}$ ,con il valore individuato nella seconda equazione, si ottiene:

$$n_h = \frac{W_h S_h}{\frac{1}{n} \sum_{h=1}^{L} W_h S_h} = n \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}$$

secondo la quale $n_h$ è proporzionale sia alla *numerosità dello strato* $W_h$

sia alla *variabilità dello strato* $S_h$

# CC Stratificato

*Individuazione degli $n_h$*
*Minimizzazione della varianza fissando i costi dell'indagine* ***C***

Nella realtà anche i costi influiscono nella determinazione degli $n_h$ ottimi. Infatti se si impone il vincolo sui costi, la conseguenza potrà essere che strati con unità costose subiranno una riduzione di $n_h$

In questo caso, facendo ricorso al medesimo procedimento, si minimizza la varianza rispetto al vincolo del costo.

$V(\bar{y}_{ST})$ minima con il vincolo che $C = C_0 + \sum_{h=1}^{L} C_h n_h$

Anche in questo caso è presente solo un vincolo, sostituendo nella precedente formula

$F_0$ $F_1$ e λ con le quantità corrispondenti cioè:

$$F_0 = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2 \qquad \lambda_1 F_1 = \lambda \left( \sum_{h=1}^{L} C_h n_h - C + C_0 \right)$$

$$F = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2 + \lambda \left( \sum_{h=1}^{L} C_h n_h - C + C_0 \right)$$

# CC Stratificato

Per trovare il minimo di tale funzione, bisogna calcolare le derivate parziali rispetto a $n_h$ e a λ e porle uguali a zero. Si ottiene il seguente sistema:

$$\frac{\partial F}{\partial n_h} = -\frac{W_h^2 S_h^2}{n_h^2} + \lambda C_h = 0 \Rightarrow \lambda C_h = \frac{W_h^2 S_h^2}{n_h^2} \Rightarrow \lambda n_h^2 = \frac{W_h^2 S_h^2}{C_h}$$

e

$$\frac{\partial F}{\partial \lambda} = \sum_{h=1}^{L} C_h n_h - C + C_0 = 0$$

Le cui soluzioni forniscono la seguente: $n_h \sqrt{\lambda} = \frac{W_h S_h}{\sqrt{C_h}}$

se sommiamo per tutti gli strati $\sum_{h=1}^{L} n_h \sqrt{\lambda} = \sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{C_h}} \quad \Rightarrow n\sqrt{\lambda} = \sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{C_h}}$

da cui: $\frac{n_h \sqrt{\lambda}}{n\sqrt{\lambda}} = \frac{\frac{W_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{C_h}}} \Rightarrow \mathrm{n_h} = \mathrm{n} \frac{\frac{W_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{C_h}}}$

Se ne deduce che i costi influiscono, con una relazione inversa, sulle dimensioni degli strati e quindi sulla dimensiona campionaria:

***a costi elevati corrispondono pochi elementi negli strati del campione***

# CC Stratificato

*Individuazione degli $n_h$*
*Minimizzazione del costo per una varianza fissata*

Nella realtà anche i costi influiscono nella determinazione degli $n_h$ ottimi. Infatti se si impone il vincolo sui costi, la conseguenza potrà essere che strati con unità costose subiranno una riduzione di $n_h$

In questo caso, facendo ricorso al medesimo procedimento, si minimizza il costo rispetto ad una varianza fissata.

$$C = C_0 + \sum_{h=1}^{L} C_h n_h \qquad \text{minima con il vincolo che} \qquad V(\bar{y}_{ST}) = V'(\bar{y}_{ST})$$

Anche in questo caso è presente solo un vincolo, sostituendo nella precedente formula

$F_0$ $F_1$ e λ con le quantità corrispondenti cioè:

$$\lambda_1 F_1 = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2 \qquad \lambda = \left( \sum_{h=1}^{L} C_h n_h - C + C_0 \right)$$

$$F = \sum_{h=1}^{L} C_h n_h - C + C_0 + \lambda \left( \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2 \right)$$

# CC Stratificato

Per trovare il minimo di tale funzione, bisogna calcolare le derivate parziali rispetto a $n_h$ e a λ e porle uguali a zero. Si ottiene il seguente sistema:

$$\frac{\partial F}{\partial n_h} = C_h + \lambda\left(-\frac{W_h^2 S_h^2}{n_h^2}\right) = 0$$

e

$$\frac{\partial F}{\partial \lambda} = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \frac{\sum W_h S_h^2}{N} - V' = 0$$

La soluzione della prima equazione è: $n_h^2 = \lambda \frac{W_h^2 S_h^2}{c_h} \quad \Rightarrow \quad n_h = \sqrt{\lambda}\frac{W_h S_h}{\sqrt{c_h}}$

che, sostituito nella seconda, comporta:

$$\sum \frac{W_h^2 S_h^2}{\sqrt{\lambda}\frac{W_h S_h}{\sqrt{c_h}}} = \frac{\sum W_h S_h^2}{N} + V' \quad \Rightarrow \quad \frac{1}{\sqrt{\lambda}} \sum W_h S_h \sqrt{c_h} = \frac{\sum W_h S_h^2}{N} + V' \quad \Rightarrow$$

$$\Rightarrow \sqrt{\lambda} = \frac{\sum W_h S_h \sqrt{c_h}}{\frac{\sum W_h S_h^2}{N} + V'}.$$

da cui:

$$n_h = \frac{\left(\sum W_h S_h \sqrt{c_h}\right)}{\frac{\sum W_h S_h^2}{N} + V'} \frac{W_h S_h}{\sqrt{c_h}}$$

# CC Stratificato

**Riflessioni sulla dimensione campionaria per strato $n_h$**

Riepilogando:

in assenza di altre informazioni, in ogni strato prendiamo un campione grande se:

A - lo strato è grande ( $N_h$ numeroso e quindi anche $W_h$ )

B - lo strato è molto variabile ( $S_h$ ampio)

C - selezionare le unità (eseguire il campionamento) in quello strato è poco costoso ( $C_h$ contenuto)

importante!!!

## In definitiva:

- l'allocazione ottima consente di conseguire una riduzione dell'errore standard superiore al 10% solo se il Coefficiente di Variazione di S è non inferiore al 50%

$$C(S)=\frac{\sqrt{\sum_{h=1}^{L}W_h\left(S_h-\bar{S}\right)^2}}{E\left(\sum_{h=1}^{L}W_hS_h\right)}$$

- la varianza delle stime si riduce quando gli $S_h$ sono notevolmente diversi fra gli strati

# CC Stratificato

## **Stima della dimensione campionaria n**

Il valore di n varia a seconda che il campione sia scelto sulla base di un valore prefissato del costo o della varianza.
Nel primo caso, ipotizzando un prefissato costo in ogni strato, occorre procedere sostituendo i rispettivi valori di $n_h$ individuati precedentemente, cioè

$$n_h = n \frac{\frac{W_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{C_h}}}$$

nella funzione di costo, si avrà: $$C - C_0 = \sum_{h=1}^{L} n_h C_h = \sum_{h=1}^{L} C_h n \frac{\frac{W_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{C_h}}}$$

e risolvendo rispetto a n si ottiene: $$n = \frac{(C - C_0) \sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} N_h S_h \sqrt{C_h}}$$

# CC Stratificato

## Stima della dimensione campionaria n

Considerando di volere **vincolare la precisione** del risultato campionario, si considera la varianza fissata ad un valore minimo e si sostituiscono gli $n_h$ ottenuti precedentemente:

$$n_h = n \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}$$

si ottiene:

$$V_{\min}\left(\bar{y}_{ST}\right) = \frac{\sum_{h=1}^{L} W_h^2 S_h^2}{n \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}} - \frac{\sum_{h=1}^{L} W_h S_h^2}{N} = \frac{\sum_{h=1}^{L} W_h S_h}{\frac{n}{\sum_{h=1}^{L} W_h S_h}} - \frac{\sum_{h=1}^{L} \frac{N_h^2}{N^2} S_h^2}{N} =$$

semplificando si avrà: $$= \frac{\sum_{h=1}^{L}\left(W_h S_h\right)\sum_{h=1}^{L}\left(W_h S_h\right)}{n} - \frac{\sum_{h=1}^{L} W_h S_h^2}{N} = \frac{\left[\sum_{h=1}^{L}\left(W_h S_h\right)\right]^2}{n} - \frac{\sum_{h=1}^{L} W_h S_h^2}{N}$$

e n sarà uguale a: $$n = \frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{V_{\min}\left(\bar{y}_{ST}\right) + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}$$

# CC Stratificato

## Casi particolari per la dimensione campionaria n

Se le varianze all'interno degli strati della popolazione $S_h^2$ , sono ignote, è possibile stimarle con $s_h^2$

$s_h$ è quindi una stima di $S_h$

In questo caso la dimensione campionaria sarà in funzione anche di $w_h$ , ovvero del peso dello strato, che ***viene fissato, a priori dal ricercatore***.
Per una varianza minima fissata, la dimensione del campione è pari a:

$$n = \frac{\sum_{h=1}^{l} \frac{W_h^2 s_h^2}{w_h}}{V\left(\bar{y}_{ST}\right) + \frac{\sum_{h=1}^{L} W_h s_h^2}{N}}$$

se possiamo ignorare la correzione per popolazione finita, possiamo riferirci alla dimensione campionaria di prima approssimazione:

$$n_0 = \frac{1}{V\left(\bar{y}_{ST}\right)} \sum_{h=1}^{l} \frac{W_h^2 s_h^2}{w_h}$$

Se $\frac{n_0}{N} > 0{,}05$ la dimensione campionaria sarà: $n = \frac{n_0}{1 + \frac{1}{NV} \sum_{h=1}^{L} W_h s_h^2}$

# CC Stratificato

## Casi particolari per la dimensione campionaria n

Se $w_h$ è scelto secondo una *allocazione proporzionale* per cui:

$$w_h = W_h = \frac{N_h}{N}$$

le due dimensioni campionarie, quella di prima approssimazione e quella effettiva, saranno rispettivamente:

$$n_0 = \frac{\sum_{h=1}^{L} W_h s_h^2}{V\left(\bar{y}_{ST}\right)} \qquad n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Se $w_h$ è scelto secondo una *allocazione ottima* per cui:

$$w_h \propto W_h S_h$$

la dimensione campionaria sarà data dalla:

$$n = \frac{\left(\sum_{h=1}^{L} W_h s_h\right)^2}{V\left(\bar{y}_{ST}\right) + \frac{1}{N}\sum_{h=1}^{L} W_h s_h^2}$$

# CC Stratificato – confronti

Possibili confronti tra il piano di campionamento casuale semplice e lo stratificato (a seconda delle diverse allocazioni) per la stima della media a parità di dimensione campionaria.

*Primo confronto*: **campionamento casuale semplice con lo stratificato, nell'allocazione proporzionale**

La varianza della media campionaria nel campione casuale semplice è pari a:

$$V(\bar{y}_{CCS}) = \frac{S^2}{n}(1-f)$$

mentre quella del campionamento casuale stratificato con allocazione proporzionale è pari a:

$$V(\bar{y}_{STprop}) = \frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2 = \frac{1}{n}\sum_{h=1}^{L} W_h S_h^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

Seguendo il criterio della scomposizione della varianza in due componenti (entro e fra i gruppi, in questo caso entro e fra gli strati) possiamo scrivere:

$$S^2 = \sum_{h=1}^{L} W_h S_h^2 + \sum_{h=1}^{L} W_h \left(\bar{Y}_h - \bar{Y}\right)^2$$

dove il primo termine dopo l'uguaglianza esprime la variabilità entro gli strati e il secondo la variabilità fra gli strati.

# CC Stratificato – confronti

Sostituendo $S^2$ nella varianza della media nel campionamento casuale semplice, si avrà:

$$V(\bar{y}_{CCS}) = \frac{(1-f)}{n} S^2 = \frac{(1-f)}{n}\left[\sum_{h=1}^{L} W_h S_h^2 + \sum_{h=1}^{L} W_h \left(\bar{Y}_h - \bar{Y}\right)^2\right] =$$

$$= \frac{(1-f)}{n}\sum_{h=1}^{L} W_h S_h^2 + \frac{(1-f)}{n}\sum_{h=1}^{L} W_h \left(\bar{Y}_h - \bar{Y}\right)^2 = V\left(\bar{y}_{STprop}\right) + \frac{(1-f)}{n}\sum_{h=1}^{L} W_h \left(\bar{Y}_h - \bar{Y}\right)^2$$

Dato che $\frac{(1-f)}{n}\sum_{h=1}^{L} W_h \left(\bar{Y}_h - \bar{Y}\right)^2$ è sempre una quantità non negativa se ne deduce che:

$$V(\bar{y}_{STprop}) \leq V(\bar{y}_{CCS})$$

L'uguaglianza si verifica se: $\frac{(1-f)}{n}\sum_{h=1}^{L} W_h \left(\bar{Y}_h - \bar{Y}\right)^2 = 0$

gli strati hanno medie fra loro uguali che, ovviamente, coincidono con la media generale.

# CC Stratificato – confronti

Possiamo calcolare il deff ottenuto dividendo la varianza dello stimatore media ottenuta con il campionamento stratificato proporzionale per la varianza della media ottenuta nel caso di campionamento casuale semplice

per ottenere questa prendi le formula originali

devi usare la proprietà per scomporre x / (a + b)

in (x / a) - b*x / (a*(a+b))

$$deff = \frac{V(\bar{y}_{STprop})}{V(\bar{y}_{CCS})} \cong 1 - \frac{\sum_{h=1}^{L} W_h \left(\bar{Y}_h - \bar{Y}\right)^2}{S^2}$$

se non ci sono differenze tra le medie di strato allora questo è uguale ad 1, maggiori sono le differenze maggiore sarà l'ultimo termine, quindi minore sarà questo deff.

Il guadagno relativo dovuto all'allocazione proporzionale si misura con:

$$\frac{V(\bar{y}_{CCS}) - V(\bar{y}_{STprop})}{V(\bar{y}_{CCS})} \cong \frac{\sum_{h=1}^{L} W_h \left(\bar{Y}_h - \bar{Y}\right)^2}{S^2}$$

# CC Stratificato – confronti

*Secondo confronto*: **campionamento casuale semplice con lo stratificato, nell'allocazione uguale**

La varianza della media campionaria nel campione casuale semplice è pari a:

$$V\left(\bar{y}_{CCS}\right)=\frac{S^2}{n}\left(1-f\right)$$

mentre quella del campionamento casuale stratificato con allocazione uguale è pari a:

$$V(\bar{y}_{STug})=\frac{L}{n}\sum_{h=1}^{L}W_h^2 S_h^2-\frac{1}{N}\sum_{h=1}^{L}W_h\, S_h^2$$

Seguendo il criterio della scomposizione della varianza in due componenti (entro e fra i gruppi, in questo caso entro e fra gli strati) possiamo scrivere la varianza del campionamento casuale semplice:

$$V\left(\bar{y}_{CCS}\right)=\frac{\left(1-f\right)}{n}S^2=\frac{\left(1-f\right)}{n}\left[\sum_{h=1}^{L}W_h S_h^2+\sum_{h=1}^{L}W_h\left(\bar{Y}_h-\bar{Y}\right)^2\right]=$$

$$=\frac{1}{n}\sum_{h=1}^{L}W_h S_h^2-\frac{1}{N}\sum_{h=1}^{L}W_h S^2{}_h+\frac{\left(1-f\right)}{n}\sum_{h=1}^{L}W_h\left(\bar{Y}_h-\bar{Y}\right)^2$$

$$=\frac{1}{n}\sum_{h=1}^{L}W_h S_h^2+V\left(\bar{y}_{STug}\right)-\frac{L}{n}\sum_{h=1}^{L}W_h^2 S_h^2+\frac{\left(1-f\right)}{n}\sum_{h=1}^{L}W_h\left(\bar{Y}_h-\bar{Y}\right)^2$$

# CC Stratificato – confronti

**Il campionamento casuale semplice fornisce risultati peggiori anche rispetto al campionamento casuale stratificato con allocazione uguale**

-perché la varianza della media è uguale a quella ottenuta con il campionamento stratificato uguale sommando anche altre quantità sicuramente non negative.

A volte, però, le due varianze non sono molto diverse fra loro.

Il **guadagno maggiore nell'adottare un piano di campionamento casuale stratificato con allocazione uguale** si ha quando si procede con **una stratificazione multipla** (con più di due variabili di stratificazione).

Inoltre, si preferisce il campionamento stratificato con allocazione uguale al campionamento casuale semplice **quando il campione è piccolo** perché si evita di prendere troppi elementi da un solo strato.

$$deff = \frac{V(\bar{y}_{STug})}{V(\bar{y}_{CCS})}$$

# CC Stratificato – confronti

*Terzo confronto*: **campionamento casuale semplice con lo stratificato, nell'allocazione ottima**

La varianza della media campionaria nel campione casuale semplice è pari a:

$$V\left(\bar{y}_{CCS}\right)=\frac{S^2}{n}\left(1-f\right)$$

mentre quella del campionamento casuale stratificato con allocazione ottima è pari a:

$$V(\bar{y}_{STott})=\frac{1}{n}\left(\sum_{h=1}^{L}W_hS_h\right)^2-\frac{1}{N}\sum_{h=1}^{L}W_hS_h^2$$

Seguendo il criterio della scomposizione della varianza in due componenti (entro e fra i gruppi, in questo caso strati) possiamo scrivere la varianza del campionamento casuale semplice:

$$V\left(\bar{y}_{CCS}\right)=\frac{1}{n}\sum_{h=1}^{L}W_hS_h^2-\frac{1}{N}\sum_{h=1}^{L}W_hS_h^2+\frac{\left(1-f\right)}{n}\sum_{h=1}^{L}W_h\left(\bar{Y}_h-\bar{Y}\right)^2$$

# CC Stratificato – confronti

*Terzo confronto*: **campionamento casuale semplice con lo stratificato, nell'allocazione ottima**

$$\frac{1}{n}\left[\sum_{h=1}^{L} W_h S_h^2 - \left(\sum_{h=1}^{L} W_h S_h\right)^2\right]$$

Il primo termine oltre l'uguaglianza rappresenta la media quadratica delle varianze $S_h$ negli strati mentre il secondo termine è la media aritmetica delle $S_h$ al quadrato, quindi la possiamo scrivere come:

$$= \frac{1}{n}\left[\sum_{h=1}^{L} W_h \left(S_h - \bar{S}\right)^2\right]$$

$$V(\bar{y}_{CCS}) = \frac{1}{n}\sum_{h=1}^{L} W_h S_h^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 + \frac{(1-f)}{n}\sum_{h=1}^{L} W_h \left(\bar{Y}_h - \bar{Y}\right)^2$$

Per cui:

$$V(\bar{y}_{CCS}) = V(\bar{y}_{STott}) + \frac{1}{n}\sum_{h=1}^{L} W_h \left(S_h - \bar{S}\right)^2 + \frac{(1-f)}{n}\sum_{h=1}^{L} W_h \left(\bar{Y}_h - \bar{Y}\right)^2$$



sono uguali quando non ci sono differenze tra le medie di strato e le varianze di strato

# CC Stratificato – confronti

*Terzo confronto*: **campionamento casuale semplice con lo stratificato, nell'allocazione ottima**

dove $$\overline{S} = \sum_{h=1}^{L} W_h S_h$$

Se ne deduce che: $V(\bar{y}_{STott}) \le V(\bar{y}_{CCS})$

$$deff = \frac{V(\bar{y}_{STott})}{V(\bar{y}_{CCS})}$$

# CC Stratificato – confronti

Possibili confronti del piano di campionamento stratificato a seconda delle diverse allocazioni) per la stima della media a parità di dimensione campionaria.

## *Quarto confronto*: **stratificato nell'allocazione proporzionale e ottima**

La varianza del campionamento casuale stratificato con *allocazione proporzionale* è pari a:

$$V(\bar{y}_{STprop}) = \frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2 = \frac{1}{n}\sum_{h=1}^{L} W_h S_h^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

La varianza del campionamento casuale stratificato con *allocazione ottima* è pari a:

$$V(\bar{y}_{STott}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

Si procede effettuando una differenza fra le due:

$$V(\bar{y}_{STprop}) - V(\bar{y}_{STott}) = \frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2 - \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

$$= \frac{1}{n}\sum_{h=1}^{L} W_h S_h^2 - \frac{\frac{n}{N}}{n}\sum_{h=1}^{L} W_h S_h^2 - \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

# CC Stratificato – confronti

Il secondo e il quarto termine de:

$$=\frac{1}{n}\sum_{h=1}^{L}W_hS_h^2-\frac{\frac{n}{N}}{n}\sum_{h=1}^{L}W_hS_h^2-\frac{1}{n}\left(\sum_{h=1}^{L}W_hS_h\right)^2+\frac{1}{N}\sum_{h=1}^{L}W_hS_h^2$$

sono uguali ma con segno opposto e quindi si possono elidere, ottenendo così:

$$=\frac{1}{n}\sum_{h=1}^{L}W_hS_h^2-\frac{1}{n}\left(\sum_{h=1}^{L}W_hS_h\right)^2=\frac{1}{n}\left[\sum_{h=1}^{L}W_hS_h^2-\left(\sum_{h=1}^{L}W_hS_h\right)^2\right]$$

Il primo termine oltre l'uguaglianza rappresenta la media quadratica delle varianze $S_h$ negli strati

mentre il secondo termine è la media aritmetica delle $S_h$ al quadrato, quindi la possiamo scrivere come:

$$=\frac{1}{n}\left[\sum_{h=1}^{L}W_h\left(S_h-\bar{S}\right)^2\right]\geq 0$$

Se ne deduce quindi che:

$$V(\bar{y}_{STprop})\geq V(\bar{y}_{STott})$$



coincidono se non ci sono differenze tra le varianze di strato

# CC Stratificato – confronti

L'uguaglianza fra le due allocazioni, si verifica quando tutti gli strati hanno la stessa varianza.

Nel caso in cui il campionamento stratificato con allocazione ottima consideri anche i costi, si deve anche verificare l'uguaglianza dei costi per ogni strato.

**L'efficienza** del piano di campionamento casuale stratificato con allocazione ottima, rispetto a quello con allocazione stratificato, si determina nel modo seguente:

$$\frac{V(\bar{y}_{STprop})}{V(\bar{y}_{STott})} = \frac{\frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2}{\frac{\sum_{h=1}^{L}(W_h S_h)^2}{n} - \frac{\sum_{h=1}^{L} W_h S_h^2}{N}} = \frac{\sum_{h=1}^{L} W_h S_h^2 - \frac{n}{N}\sum_{h=1}^{L} W_h S_h^2}{\frac{\sum_{h=1}^{L}(W_h S_h)^2}{n} - n\frac{\sum_{h=1}^{L} W_h S_h^2}{N}} \cong 1 + C^2(S)$$

dove $C(S)$ è il coefficiente di variazione di $S_h$

$$C(S) = \frac{\sqrt{\sum_{h=1}^{L} W_h \left(S_h - \bar{S}\right)^2}}{E\left(\sum_{h=1}^{L} W_h S_h\right)}$$

se coef of var è maggiore di 50% allora allocazione ottima è molto migliore altrimenti possiamo lasciar perdere

Se il coefficiente di variazione non è inferiore al 50%, si dimostra che **l'allocazione ottima migliora la precisione della stima**, in quanto l'errore quadratico medio della media è circa il 10% in meno di quello ottenuto con l'allocazione proporzionale.

# CC Stratificato – confronti

*Quinto confronto*: **stratificato nell'allocazione proporzionale e uguale**

La varianza del campionamento casuale stratificato con allocazione *proporzionale* è pari a:

$$V(\bar{y}_{STprop}) = \frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2 = \frac{1}{n}\sum_{h=1}^{L} W_h S_h^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

La varianza del campionamento casuale stratificato con allocazione *uguale* è pari a:

$$V(\bar{y}_{STug}) = \frac{L}{n}\sum_{h=1}^{L} W_h^2 S_h^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

Si procede effettuando una differenza fra le due:

$$V(\bar{y}_{STprop}) - V(\bar{y}_{STug}) = \frac{1}{n}\sum_{h=1}^{L} W_h S_h^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 - \frac{L}{n}\sum_{h=1}^{L} W_h^2 S_h^2 + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

$$= \frac{1}{n}\sum_{h=1}^{L} W_h S_h^2 - \frac{L}{n}\sum_{h=1}^{L} W_h^2 S_h^2$$

***non possiamo esprimere a priori un giudizio sulle due allocazioni***

# CC Stratificato – confronti

*Sesto confronto*: **stratificato nell'allocazione ottima e uguale**

La varianza del campionamento casuale stratificato con allocazione *ottima* è pari a:

$$V(\bar{y}_{STott}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

La varianza del campionamento casuale stratificato con allocazione *uguale* è pari a:

$$V(\bar{y}_{STug}) = \frac{L}{n}\sum_{h=1}^{L} W_h^2 S_h^2 - \frac{1}{N}\sum_{h=1}^{L} W_h\, S_h^2$$

Si procede effettuando una differenza fra le due:

$$V(\bar{y}_{STug}) - V(\bar{y}_{STott}) = \frac{L}{n}\sum_{h=1}^{L} W_h^2 S_h^2 - \frac{1}{N}\sum_{h=1}^{L} W_h\, S_h^2 - \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

$$= \frac{L}{n}\sum_{h=1}^{L} W_h^2 S_h^2 - \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2$$

***A parità di unità estratte il campionamento con allocazione ottima porta a stime più precise di quello con allocazione uguale***

# CC Stratificato – confronti

Sulla base dei confronti effettuati, si deduce che

**a) fra il campionamento casuale semplice e lo stratificato**, indipendentemente dall'allocazione adottata, **è preferibile il secondo.**

b)Ciò è sempre vero se la scelta della variabile di stratificazione è effettuata con accortezza.

Quindi *fare una stratificazione aiuta perché conduce a stime più precise*. In più, conviene prendere un numero ottimo di unità in quanto si verifica che

$$V(\bar{y}_{CCS}) \geq V(\bar{y}_{STott})$$

$$V(\bar{y}_{STprop}) \geq V(\bar{y}_{STott})$$

$$V(\bar{y}_{STug}) \geq V(\bar{y}_{STott})$$

***Si ricordi che il campionamento stratificato con allocazione uguale, rispetto a quello con allocazione proporzionale, fornisce risultati incerti.***

# CC Stratificato

**Stima dei parametri**

*Totale*

Se l'interesse è rivolto al totale della variabile Y in una popolazione generica suddivisa in L strati, si procede come nel ccs, a partire dalla media. Lo stimatore del totale di una popolazione è:

$$\hat{y}_{ST} = N\bar{y}_{ST}$$

Nel campionamento casuale stratificato, in maniera analoga alla stima della media, *la varianza dello stimatore varia in base al tipo di allocazione.*

# CC Stratificato

## *La varianza del Totale*

**Indipendentemente dal tipo di allocazione**

la varianza dello stimatore totale assume la seguente generica forma:

$$V(\hat{y}_{ST}) = \sum_{h=1}^{L} N_h (N_h - n_h) \frac{S_h^2}{n_h}$$

Infatti

$$V(\hat{y}_{ST}) = V(N\bar{y}_{ST}) = N^2 V(\bar{y}_{ST}) = N^2 \sum_{h=1}^{L} W_h^2 V(\bar{y}_h) =$$

$$= N^2 \frac{1}{N^2} \sum_{h=1}^{L} N_h^2 \frac{S_h^2}{n_h} \left( \frac{N_h - n_h}{N_h} \right) = \sum_{h=1}^{L} N_h (N_h - n_h) \frac{S_h^2}{n_h}$$

# CC Stratificato

## *La varianza del Totale*

**Allocazione uguale**

$$n_h = \frac{n}{L} \qquad \left(n_1 = n_2 = \ldots\ldots = n_L\right)$$

la varianza dello stimatore totale possiamo indicarla nel seguente modo:

$$V\left(\hat{y}_{ST}\right) = \sum_{h=1}^{L} N_h \left(N_h - n_h\right)\frac{S_h^2}{n_h} = \frac{L}{n}\sum_{h=1}^{L}\left(N_h S_h\right)^2 - \sum_{h=1}^{L} N_h S_h{}^2$$

# CC Stratificato

## *La varianza del Totale*

**Allocazione proporzionale**

$$\left(\frac{n_1}{N_1}=\frac{n_2}{N_2}=....=\frac{n_h}{N_h}=\frac{n}{N}\right)$$

la varianza dello stimatore totale possiamo indicarla nel seguente modo:

$$V\left(\hat{y}_{STprop}\right)=\frac{N-n}{n}\sum_{h=1}^{L}N_hS_h^2$$

Infatti

$$V\left(\hat{y}_{STprop}\right)=V\left(N\bar{y}_{STprop}\right)=N^2V\left(\bar{y}_{STprop}\right)=N^2\sum_{h=1}^{L}W_h^2V\left(\bar{y}_h\right)=$$

$$V\left(\hat{y}_{STprop}\right)=N^2\frac{1}{N^2}\sum_{h=1}^{L}N_h^2\frac{S_h^2}{n_h}\left(\frac{N_h-n_h}{N_h}\right)$$

# CC Stratificato

## *La varianza del Totale*

**Allocazione proporzionale**

Considerando che: $$\frac{n_1}{N_1}=\frac{n_2}{N_2}=....=\frac{n_h}{N_h}=\frac{n}{N} \qquad n_h=n\frac{N_h}{N}$$

$$V(\hat{y}_{STprop})=\sum_{h=1}^{L}N_h^2\frac{S_h^2}{\frac{nN_h}{N}}\left(\frac{N_h-\frac{nN_h}{N}}{N_h}\right)=\sum_{h=1}^{L}\frac{S_h^2}{n}\left(\frac{NN_h-\frac{NnN_h}{N}}{NN_h}\right)=$$

$$=\sum_{h=1}^{L}N_h^2\frac{S_h^2}{\frac{nN_h}{N}}\left(\frac{N_h-\frac{nN_h}{N}}{N_h}\right)=\sum_{h=1}^{L}\frac{S_h^2}{n}\left(\frac{NN_h-\frac{NnN_h}{N}}{NN_h}\right)=$$

$$=\sum_{h=1}^{L}N_h(N-n)\frac{S_h^2}{n}=\frac{N-n}{n}\sum_{h=1}^{L}S_h^2N_h$$

# CC Stratificato

## *La varianza del Totale*

**Allocazione ottima**

per ottenere la varianza dello stimatore è sufficiente sostituire alla formula generale della varianza

$$V(\hat{y}_{ST}) = \sum_{h=1}^{L} N_h (N_h - n_h) \frac{S_h^2}{n_h}$$

gli $n_h$ ottimi

Un altro modo per individuare la varianza del totale nell'allocazione ottima, consiste nel considerare la varianza della media:

$$V(\bar{y}_{STott}) = \frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{n} - \frac{\sum_{h=1}^{L} W_h^2 S_h^2}{N_h}$$

e ricavare la varianza del totale da quella della media, secondo la seguente relazione:

$$V(\hat{y}_{STott}) = V(N\bar{y}_{STott}) = N^2 V(\bar{y}_{STott})$$

# CC Stratificato

## *La varianza del Totale*

**Allocazione ottima**

si avrà:

$$V(\hat{y}_{STott}) = N^2 \frac{\frac{1}{N^2}\left(\sum_{h=1}^{L} N_h S_h\right)^2}{n} - \frac{N^2 \sum_{h=1}^{L} \frac{N_h^2}{N^2} S_h^2}{N_h} = \frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{n} - N\sum_{h=1}^{L} W_h S_h^2$$

$$= \sum_{h=1}^{L} N_h^2 \frac{S_h^2}{n_h} - \sum_{h=1}^{L} N_h S_h^2 = \frac{\left(\sum_{h=1}^{L} N_h S_h\right)^2}{n} - N\sum_{h=1}^{L} W_h S_h^2$$

Possiamo tenere conto anche dei costi. In questo caso:

$$V(\hat{y}_{STott}) = \frac{\left(\sum_{h=1}^{L} N_h S_h \sqrt{C_h}\right)^2}{C - C_0} - N\sum_{h=1}^{L} W_h S_h^2$$

dove $N\sum_{h=1}^{L} W_h S_h^2 = \sum_{h=1}^{L} N_h S_h^2$

# CC Stratificato

## Stima della dimensione campionaria n

Nel caso di stima del totale per **l'allocazione uguale**, la dimensione campionaria sarà pari a:

$$n = \frac{\displaystyle\sum_{h=1}^{L} \frac{N_h^2 s_h^2}{w_h}}{V + \displaystyle\sum_{h=1}^{L} N_h s_h^2}$$

Dove V è la varianza per un livello di precisione fissato

# CC Stratificato

**Stima della dimensione campionaria n**

Nel caso di stima del totale per **l'allocazione proporzionale**, si procede all'individuazione della dimensione campionaria di prima approssimazione

$$n_0 = \frac{N}{V}\sum_{h=1}^{L} N_h s_h^2$$

se $\frac{n_0}{N} > 0{,}05$ allora si deve determinare n: $$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Nel caso di stima del totale per **l'allocazione ottima** con varianza fissata, si avrà

$$n = \frac{\left(\sum_{h=1}^{L} N_h s_h\right)^2}{V' + \sum_{h=1}^{L} N_h s_h^2}$$

# CC Stratificato – confronti

*Primo confronto*: **campionamento casuale semplice con lo stratificato, nell'allocazione proporzionale**

La varianza del totale nel campione casuale semplice è pari a:

$$V\left(\hat{y}_{CCS}\right) = N^2 \frac{1-f}{n} S^2$$

mentre quella del campionamento casuale stratificato con allocazione proporzionale è pari a:

$$V\left(\hat{y}_{STprop}\right) = \frac{N-n}{n} \sum_{h=1}^{L} S_h^2 N_h = N^2 \frac{1-f}{n} \sum_{h=1}^{L} W_h S_h^2$$

Seguendo il criterio della scomposizione della varianza in due componenti (entro e fra i gruppi, in questo caso strati) possiamo scrivere:

$$S^2 = \sum_{h=1}^{L} W_h S_h^2 + \sum_{h=1}^{L} W_h \left(\overline{Y}_h - \overline{Y}\right)^2$$

dove il primo termine dopo l'uguaglianza esprime la variabilità entro gli strati e il secondo la variabilità fra gli strati.

# CC Stratificato – confronti

Sostituendo $S^2$ nella varianza della media nel campionamento casuale semplice, si avrà:

$$V(\bar{y}_{CCS}) = \frac{(1-f)}{n} S^2 = \frac{(1-f)}{n} \left[ \sum_{h=1}^{L} W_h S_h^2 + \sum_{h=1}^{L} W_h \left( \bar{Y}_h - \bar{Y} \right)^2 \right] =$$

ovvero:

$$V(\hat{y}_{CCS}) = V(\hat{y}_{SSTprop}) + N^2 \frac{1-f}{n} \sum_{h=1}^{L} W_h \left( \bar{Y}_h - \bar{Y} \right)^2$$

quindi

$$V(\hat{y}_{CCS}) \geq V(\hat{y}_{SSTprop})$$

come per la media campionaria, il campionamento stratificato proporzionale è più efficiente, nello stimare il totale, rispetto al campionamento casuale semplice.
Possiamo calcolare il Deff:

$$Deff = \frac{V(\hat{y}_{STprop})}{V(\hat{y}_{CCS})} \leq 1$$

# CC Stratificato – confronti

*Secondo confronto*: **campionamento casuale semplice con lo stratificato, nell'allocazione ottima**

La varianza del totale nel campione casuale semplice è pari a:

$$V(\hat{y}_{CCS}) = N^2 \frac{1-f}{n} S^2$$

mentre quella del campionamento casuale stratificato con allocazione ottima è pari a:

$$V(\hat{y}_{STott}) = \frac{1}{n}\left(N\sum_{h=1}^{L} W_h S_h\right)^2 - N\sum_{h=1}^{L} W_h S_h^2$$

Seguendo il criterio della scomposizione della varianza in due componenti (entro e fra i gruppi, in questo caso strati) possiamo scrivere la varianza del campionamento casuale semplice:

$$V(\hat{y}_{CCS}) = N^2 \frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2 + N^2 \frac{1-f}{n}\sum_{h=1}^{L} W_h \left(\bar{Y}_h - \bar{Y}\right)^2$$

$$V(\hat{y}_{CCS}) = V(\hat{y}_{STott}) + \frac{N^2}{n}\left(\sum_{h=1}^{L} W_h S_h^2 - \bar{S}^2\right) + N^2 \sum_{h=1}^{L} W_h \left(\bar{Y}_h - \bar{Y}\right)^2$$

quindi

$$V(\hat{y}_{CCS}) \geq V(\hat{y}_{STott})$$

# CC Stratificato – confronti

infatti alla varianza del campionamento con allocazione ottima vengono sommate due quantità:

$$\frac{N^2}{n}\left(\sum_{h=1}^{L} W_h S_h^2 - \bar{S}^2\right)$$

che rappresenta l'incremento della varianza imputabile al campionamento casuale semplice e

$$N^2 \sum_{h=1}^{L} W_h \left(\bar{Y}_h - \bar{Y}\right)^2$$

che invece rappresenta l'incremento della varianza imputabile all'allocazione proporzionale rispetto a quella ottima.

È possibile, anche in questo caso, calcolare l'effetto del disegno, che dato il risultato appena dimostrato sarà:

$$Deff = \frac{V(\hat{y}_{STott})}{V(\hat{y}_{CCS})} \leq 1$$

# CC Stratificato – confronti

Possibili confronti del piano di campionamento stratificato a seconda delle diverse allocazioni) per la stima del totale a parità di dimensione campionaria.

*Terzo confronto*: **stratificato nell'allocazione proporzionale e ottima**

La varianza del campionamento casuale stratificato con allocazione proporzionale è pari a:

$$V\left(\hat{y}_{STprop}\right) = N^2 \frac{1-f}{n} \sum_{h=1}^{L} W_h S_h^2 = N^2 \frac{\sum_{h=1}^{L} W_h S_h^2}{n} - N \sum_{h=1}^{L} W_h S_h^2$$

La varianza del campionamento casuale stratificato con allocazione ottima è pari a:

$$V\left(\hat{y}_{STott}\right) = \frac{1}{n}\left(N \sum_{h=1}^{L} W_h S_h\right)^2 - N \sum_{h=1}^{L} W_h S_h^2$$

Si procede effettuando una differenza fra le due:

$$V\left(\hat{y}_{STprop}\right) - V\left(\hat{y}_{STott}\right) = \frac{N^2}{n}\left(\sum_{h=1}^{L} W_h S_h^2 - \bar{S}^2\right)$$

$$V\left(\hat{y}_{STprop}\right) = V\left(\hat{y}_{STott}\right) + \frac{N^2}{n}\left(\sum_{h=1}^{L} W_h S_h^2 - \bar{S}^2\right)$$

# CC Stratificato – confronti

per cui:

$$V\left(\hat{y}_{STprop}\right) \geq V\left(\hat{y}_{STott}\right)$$

l'uguaglianza si verifica quando tutti gli strati hanno la stessa varianza.
Pertanto se negli strati le varianze sono uguali, l'allocazione proporzionale coincide con quella ottima.

Se si può trascurare il fattore di correzione per popolazione finita, un altro modo per confrontare tali varianze consiste nel considerare il coefficiente di variazione. Infatti, si può dimostrare che il rapporto:

$$\frac{V(\hat{y}_{STprop})}{V(\hat{y}_{STott})} \cong 1 + C^2(S)$$

dove: $C(S) = \dfrac{\sqrt{V(S_h)}}{E(S)}$

tale rapporto è pari a: $$\frac{V(\hat{y}_{STprop})}{V(\hat{y}_{STott})} = \frac{\sum_{h=1}^{L} W_h S_h^2}{\left(\sum_{h=1}^{L} W_h S_h\right)^2}$$

# CC Stratificato

**Stima dei parametri**

*Proporzione*

La stima di una proporzione o di una frequenza relativa si riconduce alla stima della media nel caso di variabile casuale che assume valore 0 o 1.

Come è stato fatto per il campionamento casuale semplice, si può procedere anche alla stima del numeratore della frequenza che anch'esso è riconducibile al totale di una variabile dicotoma (0-1).

$$P_h = \frac{A_h}{N_h} \quad \text{e} \quad Q_h = 1 - P_h$$

sono rispettivamente la proporzione dell'evento successo e il suo complementare nella popolazione.

$$p_h = \frac{a_h}{n_h} \quad \text{e} \quad q_h = 1 - p_h$$

sono rispettivamente la proporzione dell'evento successo e il suo complementare nel campione.

# CC Stratificato

**Stima dei parametri**

*Proporzione*

Lo stimatore della proporzione $P_{ST} = \sum_{h=1}^{L} \frac{N_h P_h}{N}$ è: $p_{ST} = \sum_{h=1}^{L} \frac{N_h p_h}{N}$

Mentre quello del numeratore $A_{ST} = \sum_{h=1}^{L} N_h P_h$ è: $a_{ST} = \sum_{h=1}^{L} N_h p_h$

*La varianza della proporzione*

$$V(p_{ST}) = \sum_{h=1}^{L} W_h^2 V(p_h) = \frac{1}{N^2} \sum_{h=1}^{L} N_h^2 \frac{(N_h - n_h)}{N_h - 1} \frac{P_h Q_h}{n_h}$$

dove $V(p_h) = \frac{P_h Q_h}{n_h} \frac{(N_h - n_h)}{N_h - 1}$ è la varianza nello strato h-esimo

$\frac{(N_h - n_h)}{N_h - 1}$ è il fattore di correzione se la popolazione è finita.

# CC Stratificato

## *La varianza della proporzione*

Quando il termine

$$\frac{1}{N_h}$$

è trascurabile in ogni strato, possiamo indicare la varianza all'interno di ogni strato nel seguente modo:

$$V(p_{ST}) = \sum_{h=1}^{L} \frac{W_h^2 P_h Q_h}{n_h}(1 - f_h)$$

Se possiamo omettere la correzione per popolazione finita, allora essa diventerà:

$$V(p_{ST}) = \sum_{h=1}^{L} W_h^2 \frac{P_h Q_h}{n_h}$$

Queste formule per la varianza dello stimatore valgono per il caso generale e per l'**allocazione uguale**.

# CC Stratificato

## *La varianza della proporzione*

**Allocazione proporzionale** $\left(\frac{n_1}{N_1}=\frac{n_2}{N_2}=....=\frac{n_h}{N_h}=\frac{n}{N}\right)$ $\quad n_h=n\frac{N_h}{N}$

la varianza dello stimatore totale possiamo indicarla nel seguente modo:

$$V(p_{STprop})=\frac{1}{N^2}\sum_{h=1}^{L}N_h^2\frac{\left(N_h-n\frac{N_h}{N}\right)}{N_h-1}\frac{P_hQ_h}{n\frac{N_h}{N}}=\frac{1}{N^2}\sum_{h=1}^{L}N_h^2\frac{(NN_h-nN_h)}{N_h-1}\frac{NP_hQ_h}{nN_h}=$$

$$=\frac{1}{N}\sum_{h=1}^{L}N_h\frac{N_h(N-n)}{N_h-1}\frac{P_hQ_h}{n}=\frac{1}{N}\sum_{h=1}^{L}N_h^2\frac{N_h(N-n)}{N(N_h-1)}\frac{P_hQ_h}{n}=$$

$$=\frac{N-n}{N}\frac{1}{Nn}\sum_{h=1}^{L}N_h^2\frac{P_hQ_h}{N_h-1}\approx\frac{1-f}{n}\sum_{h=1}^{L}N_h\frac{P_hQ_h}{N}=\frac{1-f}{n}\sum_{h=1}^{L}W_hP_hQ_h$$

per cui $\quad V(p_{STprop})=\frac{1-f}{n}\sum_{h=1}^{L}W_hP_hQ_h$

# CC Stratificato

## *La varianza della proporzione*

**Allocazione ottima**

Possiamo indicare la varianza dello stimatore in questo modo:

$$V(p_{STott}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h \sqrt{P_h Q_h}\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h P_h Q_h$$

ottenuta considerando la scelta di $n_h$ per minimizzare la varianza ovvero

$$n_h \propto N_h \sqrt{\frac{N_h}{N_h - 1}} \sqrt{P_h Q_h} \cong N_h \sqrt{P_h Q_h}$$

$\frac{1}{N}\sum_{h=1}^{L} W_h P_h Q_h$ è il fattore di correzione se la popolazione è finita.

$P_h$ e $Q_h$ nelle formule per determinare la varianza vengono sostituite rispettivamente dalle loro stime $p_h$ e $q_h$

# CC Stratificato

**Stima della dimensione campionaria n**

Prima di determinare la dimensione del campione è bene specificare che dagli

$$n_h \propto N_h \sqrt{\frac{N_h}{N_h - 1}} \sqrt{P_h Q_h} \cong N_h \sqrt{P_h Q_h}$$

si ricava che:

$$n_h = n \frac{N_h \sqrt{P_h Q_h}}{\sum_{h=1}^{L} N_h \sqrt{P_h Q_h}}$$

e ponendo la varianza minima per costo fissato si ottiene:

$$n_h = n \frac{N_h \sqrt{\frac{P_h Q_h}{C_h}}}{\sum_{h=1}^{L} N_h \sqrt{\frac{P_h Q_h}{C_h}}}$$

# CC Stratificato

**Stima della dimensione campionaria n**

Nel caso di stima della proporzione per **l'allocazione proporzionale**, si procede all'individuazione della dimensione campionaria di prima approssimazione

$$n_0 = \frac{1}{V}\sum_{h=1}^{L} W_h p_h q_h$$

se $\frac{n_0}{N} > 0{,}05$ allora si deve determinare n: $n = \dfrac{n_0}{1+\dfrac{n_0}{N}}$

Nel caso di stima della proporzione per **l'allocazione ottima**, si avrà

$$n_0 = \frac{1}{V}\left(\sum_{h=1}^{L} W_h \sqrt{p_h q_h}\right)^2 \qquad n = \frac{n_0}{1+\dfrac{1}{NV}\sum_{h=1}^{L} W_h p_h q_h}$$

tali determinazioni sono state ottenute considerando la seguente uguaglianza $\dfrac{N_h}{N_h - 1} \equiv 1$

# CC Stratificato – confronti

Per valutare l'efficienza del campionamento stratificato, con allocazione uguale e ottima, rispetto al campionamento casuale semplice, ci si attiene rispettivamente alla regola che considera più efficiente il campionamento stratificato se le proporzioni di strato

$$P_h$$

*sono sensibilmente differenti fra loro*

**Primo confronto: Campionamento casuale semplice e stratificato ottimo**

Se si indaga nello specifico sull'efficienza del *campionamento stratificato con allocazione ottima rispetto a quella del campione casuale semplice* è opportuno specificare ulteriormente che se i costi per unità sono uguali in tutti gli strati, il guadagno del campionamento stratificato rispetto al campionamento casuale semplice è modesto a meno che i

$$P_h$$

*non varino molto di strato in strato*

# CC Stratificato – confronti

Possibili confronti del piano di campionamento stratificato a seconda delle diverse allocazioni) per la stima della proporzione a parità di dimensione campionaria.

**Secondo confronto: Campionamento casuale stratificato ottimo e proporzionale**

Se, al contrario, si indaga sull'efficienza dello stratificato con allocazione ottima rispetto allo stesso con allocazione proporzionale si dimostra che il campionamento stratificato ottimo consente stime più precise se n è fissato e $P_h$ non è compreso fra 0,1 e 0,9

$$\frac{V(p_{STott})}{V(p_{STprop})} = \frac{\left(\sum_{h=1}^{L} W_h \sqrt{P_h Q_h}\right)^2}{\sum_{h=1}^{L} W_h P_h Q_h}$$

Tale rapporto sarà molto più piccolo di 1 quando $P_h \leq 0{,}1$ o $P_h \geq 0{,}9$

La precisione relativa minore si avrà se si verificano le seguenti condizioni:
- due strati di ampiezza costante $W_1 = W_2$
- $P_1 = 0{,}5$ e $P_2 = P_0$

$$D = \frac{\left(0{,}5 + \sqrt{P_0 Q_0}\right)^2}{2\left(0{,}25 + P_0 Q_0\right)}$$

in questo caso conviene utilizzare l'ottimo solo se p ha valore estremo

altrimenti meglio prop per semplicità se P ha altri valori e l'ampiezza degli strati è costante

# CC Stratificato – considerazioni

se conosciamo la variabile di stratificazione bene, allora conviene sempre stratificare per ottenere gruppi omogenei all'interno ed eterogenei tra loro

1. per tutti gli stimatori considerati, ***la stratificazione produce una soluzione auspicabile quando si conosce la distribuzione della variabile oggetto di studio***, perché essa consente di stratificare bene e quindi massimizzare la differenza fra gli strati, rendendo gli stessi, al loro interno, il più possibile omogenei.
2. In alternativa si ricorre ad informazioni supplementari con l'obiettivo di individuare dei caratteri di stratificazione in relazione nota con la variabile di indagine.
3. La stratificazione è il risultato di una serie di scelte circa il carattere o i caratteri di stratificazione, la determinazione del numero di strati, la delimitazione degli strati, l'allocazione delle unità campionarie tra gli strati, ecc. Ogni scelta deve essere ben ponderata perché può inficiare il buon esito dell'intera indagine.
4. Con la stratificazione, inoltre, possiamo scegliere opportunamente diversificando la procedura di estrazione all'interno di ogni strato. Ciò consente di tenere conto della diversa distribuzione dei gruppi di popolazione, della diversa natura delle informazioni nelle liste, ecc.

La diversificazione e la stratificazione consentono anche di stimare alcuni parametri per una parte della popolazione ovvero analizzare in modo approfondito solo alcuni domini di studio ovvero solo una parte della popolazione per cui sono pianificate nel piano campionario stime separate.