Campionamento Stratificato: Proprietà e Stimatori

- Campionamento stratificato;
- Pregi e difetti del campionamento stratificato;
- Post stratificazione
- Tecniche di allocazione;
- Stimatore ponderato del totale;
- Varianza dello stimatore con allocazione proporzionale;
- Stimatore della media;
- Stimatore della proporzione;
- Guadagno di efficienza dell'allocazione proporzionale rispetto al CCS;
- Guadagno di efficienza dell'allocazione ottimale rispetto all'allocazione proporzionale;
- Stimatore per quoziente;

Campionamento stratificato

Il campionamento stratificato è un campionamento probabilistico che consiste nella classificazione della popolazione in subpopolazioni (strati), sulla base delle informazioni ausiliarie e nella selezione di campioni indipendenti da ciascuno strato.

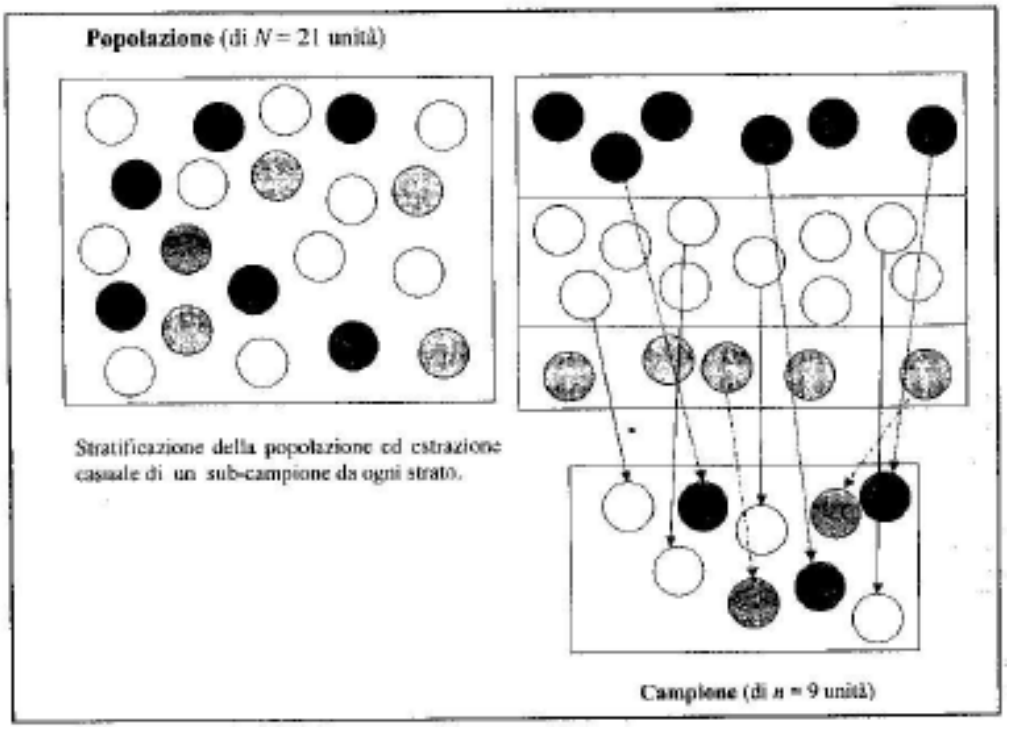
Divisione in strati (sottogruppi) della popolazione !!!

Estrazione casuale delle osservazioni presenti in ciascuno strato in maniera indipendente .

Caratteristiche:

1. Gli strati sono tali da essere OMOGENEI al loro interno ed ETEROGENEI tra loro (miglioramento delle stime);
2. Gli strati non devono sovrapporsi, ovvero ogni unità della lista deve appartenere ad uno e ad uno solo degli strati.
3. Gli strati permettono di ottenere un miglioramento delle stime a parità di numerosità campionaria oppure di contenere la numerosità campionaria a parità di efficienza.

4. Da ogni strato di estrae un campione casuale semplice; si hanno tanti campioni quanti sono gli strati. Tali campioni sono indipendenti tra loro e possono avere dimensioni campionarie differenti.



Esempio di stratificazione:

- **area geografica** – es.: regioni, città, rioni ecc.;
- **sesso**
- **età**– es.: classi di età;
- **reddito** – classi di reddito.

Perché preferire il campionamento stratificato a quello casuale semplice???

Esempio: Si consideri una classe di studenti liceali composta da 17 maschi e 13 femmine e si supponga di formare un campione di 4 unità allo scopo di stimare la statura media.

Campioni possibili: 27.405

di cui
- 2.380 composti da soli maschi
- 715 composti da sole femmine

- Il considerare in un campione solo maschi, o la maggioranza di questi, porterebbe ad una sovrastima del fenomeno.
- Il considerare in un campione solo femmine, o la maggioranza di queste, porterebbe ad una sottostima del fenomeno.

Con il campionamento stratificato si persegue il miglioramento della precisione degli stimatori dei parametri della popolazione rispetto al campionamento dall'intera popolazione di unità elementari, sfruttando la possibilità di campionare separatamente all'interno di diverse sottopopolazioni, e quindi meglio rappresentando la popolazione stessa.

La conoscenza di variabili ausiliarie in popolazione è alla base dell'adozione di strategie diverse da quelle basate sul campionamento casuale semplice e sull'impiego di stimatori diretti. Le informazioni ausiliarie possono essere usate per costituire gruppi.

Condizione necessaria per la realizzazione del campionamento stratificato è la conoscenza, per ciascun elemento della popolazione, del valore di una variabile ausiliaria per assegnare ciascun elemento della popolazione ad uno di L strati esaustivi e mutualmente esclusivi.

La suddivisione di una popolazione in strati è un esempio di raggruppamento di unità elementari in unità complesse.

La stratificazione può fornire notevoli guadagni nell'efficienza delle stime senza uscire dall'idea del campionamento casuale semplice, che continua a valere all'interno degli strati.

E' l'unico procedimento che assicura valutazioni all'interno di ciascuna sottopopolazione. Di solito non dà risultati peggiori del campionamento casuale semplice, a meno che non si verifichi un'allocazione del campione che stravolga completamente le considerazioni sulla omogeneità

all'interno degli strati e sull'importanza relativa degli strati stessi.

Se la suddivisione in strati, e quindi la conoscenza dei valori della variabile ausiliaria, non è costosa da ottenere, vale la pena di impiegare questa tecnica.

L'onerosità di questo metodo sta nell'obbligo di costruire tante liste quante sono le sottopopolazioni evidenziate.

A questo riguardo si vedrà che meno impegnativo è invece il campionamento a grappolo che segue la logica opposta a quella della stratificazione.

Pregi e difetti del campionamento stratificato

PREGI

- Garantisce una migliore rappresentatività rispetto al CCS generando quindi stime più efficienti a parità di numerosità campionaria;
- A parità di efficienza, consente di contenere la numerosità campionaria;
- Sfrutta informazioni disponibili sulla popolazione.

DIFETTI

- Richiede un lavoro preliminare alla rilevazione come la scelta delle variabili di stratificazione
- Elenco delle unità esclusivamente legato alle ripartizioni della popolazione
- Conduzione di un'indagine pilota per misurare l'efficacia del campionamento stratificato.

Esempi di indagini campionarie con stratificazione delle unità

Esempio 1: Indagine Trenitalia sulla mobilità sistematica. L'obiettivo è valutare la quota di mercato del treno rispetto agli altri mezzi di trasporto su scala nazionale. La procedura è quella si intervistare i viaggiatori e capire che mezzo di trasporto utilizzano.
Variabili di stratificazione: *??*

Esempio 2: Indagine del Ministero dell'Istruzione sugli sbocchi occupazionali: studiare su scala nazionale le tipologie di sbocco a posteriori al percorso di laurea.
Variabili di stratificazione: *??*

Esempio 3: Indagine sull'andamento della stagione turistica su scala regionale.
Variabili di stratificazione: *??*

Simbologia:

Sia

n - la dimensione del campione;

N - la dimensione della popolazione le cui osservazioni sono ordinate secondo un certo criterio;

L - numero di sottopopolazioni o di strati;

n_h - la dimensione del campione nello strato h;

N_h - la dimensione della popolazione nello strato h;

Y_{hi} - il valore della variabile Y posseduto dall'unità i della popolazione dello strato h;

y_{hi} - il valore della variabile Y posseduto dall'unità i del campione estratto dallo strato h.

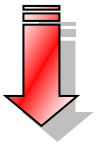
Alcune relazioni:

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

La probabilità di inclusione del primo ordine dell'unità i appartenente allo strato h è:

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

Come effettuare la stratificazione????



Valutazione dei caratteri della popolazione che potrebbero essere correlati con la variabile oggetto di indagine.

Quanto più stretto è il legame fra le variabili di stratificazione e la variabile oggetto di indagine, tanto più efficace sarà la stratificazione

---

Esempio: stima del tasso di occupazione sul territorio per una popolazione caratterizzata da N=30.

1° ipotesi: stratificazione per sesso;

2° ipotesi: stratificazione per zona di provenienza;

3° ipotesi: stratificazione per classi di età;

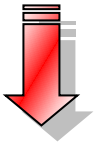
4° ipotesi: stratificazione per sesso, classi di età, zona di provenienza;

Regole di stratificazione….

- Si ottengono risultati migliori se si stratifica impiegando più di una variabile piuttosto che una variabile con numerose modalità;

- Le stime sono tendenzialmente più efficienti se le variabili di strato sono indipendenti tra loro *(rischio di riprodurre informazioni ridondanti)*;

L'efficienza aumenta con l'aumentare del numero di strati



Dopo un certo numero di strati il guadagno di efficienza diventa modesto!!!

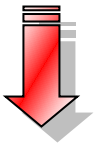
Unità autorappresentative:

$$Y^* < \mu + S\sqrt{N}\,/\,n$$

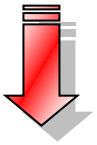
dove
-μ e S sono la media e la varianza dell'intera popolazione;
-Y* e' il valore soglia.

Tecniche di allocazione

Determinare *l'allocazione del campione* tra gli strati significa decidere quante unità, tra le n appartenenti al campione, devono essere selezionate all'interno di ciascuno strato



Tecniche di allocazione tra gli strati!!!



1. Allocazione *uniforme*;
2. Allocazione *proporzionale*;
3. Allocazione *ottimale*;

La scelta della tipologia di allocazione influisce pesantemente sui costi e sulla precisione delle stime ma non esiste una regola ottimale a priori: bisogna valutare caso per caso

Allocazione uniforme (o uguale):

In questa tipologia di allocazione, all'interno di ciascuno degli L strati verrà campionato lo stesso numero n_h di unità dando luogo alla seguente situazione:

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

N.B.: Se gli strati sono di uguale dimensione, l'allocazione uniforme coincide con l'allocazione proporzionale.

Esempio: si vuole effettuare un campionamento sugli studenti frequentanti un corso universitario per un'indagine sulla valutazione della didattica. Da una popolazione di 250 studenti (180 maschi e 70 femmine) si vuole estrarre un campione di 50 unità.

1) Stratificazione per sesso:

Strati: L_1=Maschi; L_2=Femmine

2) Allocazione uniforme:

L_1=50/2=25 ; L_2=50/2=25

Allocazione proporzionale (o autoponderante):

In questa tipologia di allocazione la numerosità campionaria è

$$n_h = \frac{nN_h}{N}$$

in modo che il *numero di unità campionate all'interno del generico strato è direttamente proporzionale all'ampiezza dello strato stesso.*

Il rapporto nasce dalla considerazione che:

$$\frac{n_h}{N_h} = \frac{n}{N}$$

N.B.: L'adozione dell'allocazione proporzionale consente di semplificare in maniera considerevole le procedure di calcolo.

Esempio: Da una popolazione di 250 studenti (180 maschi e 70 femmine) si vuole estrarre un campione di 50 unità.

1) Stratificazione per sesso:

Strati: L_1=Maschi; L_2=Femmine

2) Allocazione proporzionale:

$$n_1 = \frac{50 \cdot 180}{250} = 36 \quad ; \quad n_1 = \frac{50 \cdot 70}{250} = 14$$

Allocazione ottimale:

Nel caso si riuscisse a conoscere, come informazione aggiuntiva, la varianza del carattere all'interno di ogni strato, l'allocazione risulta

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

N.B.: Nella allocazione ottimale un maggior numero di unità verrà campionato da quegli strati che hanno una varianza interna maggiore.

Se si dispone delle varianze di strato, l'allocazione ottimale è preferibile all'allocazione proporzionale!!

Esempio: Si vuole stimare il numero medio di libri venduti in un giorno in una grande città stratificata in 3 zone distinte. Se N è il numero di librerie presenti sul territorio, Y il numero medio ed S lo s.q.m. di libri venduti in un giorno dalle librerie, estrarre un campione n=30 unità con piano di campionamento stratificato.

Strato 1: $\mathbf{N_1}$=40; $\mathbf{Y_1}$=120; $\mathbf{S_1}$=4
Strato 2: $\mathbf{N_2}$=36; $\mathbf{Y_2}$=63; $\mathbf{S_2}$=8
Strato 3: $\mathbf{N_3}$=24; $\mathbf{Y_3}$=60; $\mathbf{S_3}$=3

$$n_1 = \frac{40 \cdot 4}{520} 30 = 10 \quad ; \quad n_2 = \frac{36 \cdot 8}{520} 30 = 17 \quad ; \quad n_3 = \frac{24 \cdot 3}{520} 30 = 5$$

Stimatore ponderato del totale

Totale:

$$\hat{Y}_{ST} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{y_{hi}}{\pi_{hi}}$$

da cui

$$\hat{Y}_{ST} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{y_{hi}}{n_{hi} / N_{hi}} = \sum_{h=1}^{L} N_h \bar{y}_h$$

Lo stimatore del totale è uguale alla somma degli stimatori per espansione del totale dei singoli strati!!!

Varianza dello stimatore:

$$Var\left(\hat{Y}_{ST}\right) = Var\left(\sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{y_{hi}}{\pi_{hi}} \right) = Var\left(\sum_{h=1}^{L} N_h \bar{y}_h \right) =$$

$$= \sum_{h=1}^{L} Var\left(N_h \bar{y}_h\right) = \sum_{h=1}^{L} N_h^2 \frac{\left(1 - f_h\right)}{n_h} \cdot S_h^2$$

Stima della varianza sul campione:

$$v\left(\hat{Y}_{ST}\right) = \sum_{h=1}^{L} N_h^2 \frac{\left(1 - f_h\right)}{n_h} \cdot s_h^2$$

L'efficienza dello stimatore dipende dalla varianza entro gli strati.
L'efficienza del campionamento dipende dall'omogeneità all'interno dei gruppi.

Varianza dello stimatore con allocazione proporzionale

Si consideri l'espressione della varianza dello stimatore del totale:

$$V\left(\hat{Y}_{ST}\right)=\sum_{h=1}^{L} N_h^2 \frac{\left(1-f_h\right)}{n_h}\cdot s_h^2$$

Sapendo che

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

sostituendo:

$$Var\left(\hat{Y}_{ST}\right)=\sum_{h=1}^{L} N_h\cdot N_h \frac{\left(1-f\right)}{n_h}\cdot S_h^2=\left(1-f\right)\frac{N}{n}\sum_{h=1}^{L} N_h\cdot S_h^2$$

Moltiplicando e dividendo per N:

$$Var\left(\hat{Y}_{ST}\right)=\frac{N^2}{n}\left(1-f\right)\sum_{h=1}^{L}\underbrace{\frac{N_h}{N}}_{W_h}\cdot S_h^2$$

In definitiva:

$$Var_p\left(\hat{Y}_{ST}\right)=\frac{N^2}{n}\left(1-f\right)\sum_{h=1}^{L} W_h\cdot S_h^2$$

Nell'allocazione proporzionale la varianza dello stimatore del totale è proporzionale alla media aritmetica ponderata delle varianze degli strati!!!

Se la varianza in ciascuno strato è pressoché la stessa:

$$S_h^2 \cong \tilde{S}^2, \forall h$$

la varianza dello stimatore del totale diventa:

$$V_{\text{prop}}(\hat{Y}_{ST}) = N^2 \frac{1-f}{n} \sum_{h=1}^{H} \frac{N_h}{N} \tilde{S}^2 = N^2 \frac{1-f}{n} \tilde{S}^2 \underbrace{\sum_{h=1}^{H} \frac{N_h}{N}}_{1} = N^2 \frac{1-f}{n} \tilde{S}^2$$

L'efficienza dello stimatore del totale nel campionamento stratificato è la stessa dello stimatore nel campionamento casuale semplice

$$Var_{\text{p}}(\hat{Y}_{ST}) \cong Var(\hat{Y})$$

Confronto di efficienze con il CCS:

Esempio

Si consideri una popolazione di ampiezza N=8 così composta:

Unità	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Y	5	5	7	6	20	25	28	28

Si estragga un campione di n=4 e si calcolino le varianze degli stimatori del totale per il campionamento stratificato con allocazione uniforme e per il campionamento casuale semplice e si confrontino i risultati.

1) CCS:

$$Var(\hat{Y}) = N^2 \frac{1-f}{n} S^2 = 8 \cdot \frac{1-\frac{4}{8}}{4} \cdot 115{,}14 = 921{,}12$$

2) Campionamento stratificato:

Si divida la popolazione in due strati costituiti da 4 unità ciascuno e si estraggano due unità da ogni strato:

$$N_1 = N_2 = 4 \quad ; \quad n_1 = n_2 = 2 \quad ; \quad S_1^2 = 0{,}92 \quad ; \quad S_2^2 = 14{,}25$$

$$Var\left(\hat{Y}_{ST}\right) = \sum_{h=1}^{L} N_h^2 \frac{(1-f_h)}{n_h} \cdot S_h^2 = 4^2 \frac{1-\frac{2}{4}}{2} 0{,}92 + 4^2 \frac{1-\frac{2}{4}}{2} 14{,}25 = 60{,}68$$

$$Var\left(\hat{Y}_{ST}\right) < Var\left(\hat{Y}\right)$$

Stimatore della media

Lo stimatore della media campionaria ponderata a probabilità di inclusione costanti è:

$$\hat{\bar{Y}}_{ST} = \frac{1}{N}\hat{Y}_{ST} = \frac{1}{N}\cdot\sum_{h=1}^{L} N_h \cdot \bar{y}_h = \sum_{h=1}^{L} W_h \cdot \bar{y}_h$$

La varianza dello stimatore risulta:

$$Var\,(\hat{\bar{Y}}_{ST}) = Var\left(\frac{\hat{Y}_{ST}}{N}\right) = \frac{1}{N^2}\cdot Var\left(\hat{Y}_{ST}\right) = \frac{1-f}{n}\sum_{h=1}^{L} W_h \cdot S_h^2$$

Stima della varianza sul campione:

$$v(\hat{\bar{Y}}_{ST}) = \frac{1-f}{n}\sum_{h=1}^{L} W_h \cdot s_h^2$$

s^2: varianza campionaria.

Stimatore della proporzione

Si consideri lo stimatore della media:

$$\hat{\bar{Y}}_{ST} = \sum_{h=1}^{L} W_h \cdot \bar{y}_h$$

La cui varianza risulta:

$$Var(\hat{\bar{Y}}_{ST}) = \frac{1-f}{n} \cdot \sum_{h=1}^{L} W_h \cdot S_h^2$$

Per caratteri dicotomici (si ipotizza la distribuzione bernoulliana), lo stimatore della proporzione diventa:

$$\hat{P}_{ST} = \sum_{h=1}^{L} W_h \cdot p_h$$

con relativa varianza:

$$v(\hat{P}_{ST}) = \frac{1-f}{n} \cdot \sum_{h=1}^{L} W_h \cdot p_h(1-p_h)$$

Guadagno in efficienza dell'allocazione proporzionale rispetto al CCS

Si vuole quantificare la relazione:

$$Var\left(\hat{\bar{Y}}\right) - Var_p\left(\hat{\bar{Y}}_{ST}\right)$$

Si parta dall'espressione della scomposizione della devianza:

$$(N-1)S^2 = \underbrace{\sum_{h=1}^{L}(N_h - 1)\cdot S_h^2}_{\text{Devianza nei gruppi}} + \underbrace{\sum_{h=1}^{L} N_h \cdot \left(\bar{Y}_h - \bar{Y}\right)^2}_{\text{Devianza tra i gruppi}}$$

DIM:

$$S^2 = \frac{1}{N-1}\sum_{j=1}^{N}\left(Y_j - \bar{Y}\right)^2 \quad \Rightarrow \quad S^2(N-1) = \sum_{j=1}^{N}\left(Y_j - \bar{Y}\right)^2$$

Dividendo la popolazione in strati:

$$S^2(N-1) = \sum_{h=1}^{L}\sum_{j=1}^{N_h}\left(Y_{hj} - \bar{Y}\right)^2$$

$$S^2(N-1)=\sum_{h=1}^{L}\sum_{j=1}^{N_h}\left(Y_{hj}^2+\bar{Y}^2-2Y_{hj}\bar{Y}\right)$$

$$S^2(N-1)=\sum_{h=1}^{L}\sum_{j=1}^{N_h}Y_{hj}^2+\sum_{h=1}^{L}\sum_{j=1}^{N_h}\bar{Y}^2-\sum_{h=1}^{L}\sum_{j=1}^{N_h}2Y_{hj}\bar{Y}$$

$$S^2(N-1)=\sum_{h=1}^{L}\sum_{j=1}^{N_h}Y_{hj}^2+\bar{Y}^2\cdot N-2\bar{Y}\cdot\bar{Y}N$$

$$S^2(N-1)=\sum_{h=1}^{L}\sum_{j=1}^{N_h}Y_{hj}^2-\bar{Y}^2\cdot N$$

Aggiungendo e sottraendo il totale ponderato delle medie nell'espressione precedente:

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

$$S^2(N-1)=\sum_{h=1}^{L}\sum_{j=1}^{N_h}Y_{hj}^2-\sum_{h=1}^{L}N_h\left(\bar{Y}_h\right)^2+\sum_{h=1}^{L}N_h\left(\bar{Y}_h\right)^2-\bar{Y}^2\cdot N$$

Moltiplico per N_h / N_h

$$S^2(N-1)=\sum_{h=1}^{L}N_h\frac{1}{N_h}\sum_{j=1}^{N_h}Y_{hj}^2-\sum_{h=1}^{L}N_h\left(\bar{Y}_h\right)^2+\sum_{h=1}^{L}N_h\left(\bar{Y}_h\right)^2-\bar{Y}^2\cdot N$$

$$S^2(N-1)=\sum_{h=1}^{L}N_h\underbrace{\left[\frac{1}{N_h}\sum_{j=1}^{N_h}Y_{hj}^2-\left(\bar{Y}_h\right)^2\right]}_{=S^2_h\ (N_h-1)}+\underbrace{\sum_{h=1}^{L}N_h\left(\bar{Y}_h\right)^2-\bar{Y}^2\cdot N}_{=\Sigma\, N_h\,(\bar{Y}_h-\bar{Y})^2}$$

In definitiva:

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

c.v.d.!!!

Si procede ora alla quantificazione del guadagno di efficienza.

DIM:

Bisogna quantificare la relazione:

$$Var(\hat{Y})-Var_p(\hat{Y}_{ST})$$

Sapendo che:

$$Var(\hat{Y})=N^2\frac{1-f}{n}S^2$$

$$Var(\hat{Y}_{ST})=N^2\frac{1-f}{n}\sum_{h=1}^{L}\frac{N_h}{N}S_h^2$$

La differenza tra le varianze degli stimatori del totale è:

$$Var(\hat{Y})-Var(\hat{Y}_{ST})=N^2\frac{1-f}{n}S^2-N^2\frac{1-f}{n}\sum_{h=1}^{L}\frac{N_h}{N}S_h^2$$

$$Var(\hat{Y})-Var(\hat{Y}_{ST})=N^2\frac{1-f}{n}\left[S^2-\sum_{h=1}^{L}\frac{N_h}{N}S_h^2\right]$$

Si consideri l'espressione della scomposizione della devianza e si dividano ambo i membri per N-1:

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

Per N che tende a $+\infty$

$$= \frac{N_h - 1}{N-1} = \frac{N_h}{N} = \frac{N_h}{N-1}$$

Da cui

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

$$Var(\hat{Y}) - Var_p(\hat{Y}_{ST}) = N^2\frac{1-f}{n}\left[\sum_{h=1}^{L}\frac{N_h}{N}S_h^2 + \sum_{h=1}^{L}\frac{N_h}{N}\left(\bar{Y}_h - \bar{Y}\right)^2 - \sum_{h=1}^{L}\frac{N_h}{N}S_h^2\right]$$

In definitiva:

$$Var(\hat{Y}) = N^2\frac{1-f}{n}\sum_{h=1}^{L}\frac{N_h}{N}\left(\bar{Y}_h - \bar{Y}\right)^2 + Var_p(\hat{Y}_{ST})$$

c.v.d.!!!

Nel campionamento stratificato con allocazione proporzionale il

GUADAGNO IN EFFICIENZA

rispetto al campionamento casuale semplice è tanto maggiore quanto più le medie degli strati differiscono tra loro!!!

In definitiva è stato dimostrato che:

$$Var(\hat{Y}) \geq Var_p(\hat{Y}_{ST})$$

L'efficienza si può esprimere anche in termini di rapporto tra varianze:

$$Deff = \frac{Var_p(\hat{Y}_{ST})}{Var(\hat{Y})} = \frac{N^2 \frac{1-f}{n} \sum_{h=1}^{L} W_h \cdot S_h^2}{N^2 \frac{1-f}{n} S^2}$$

Effetto del disegno

$$Deff = \frac{Var_p(\hat{Y}_{ST})}{Var(\hat{Y})} = \frac{\sum_{h=1}^{L} W_h \cdot S_h^2}{S^2}$$

Il rapporto esprime la riduzione della varianza che consegue con l'allocazione proporzionale.

Guadagno di efficienza dell'allocazione ottimale rispetto all'allocazione proporzionale

Si consideri la varianza dello stimatore della media calcolato nel campionamento stratificato:

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

Questa relazione può essere scritta anche:

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

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

Sostituendo nella relazione precedente il valore dell'allocazione ottimale n_h:

$$Var_o(\hat{\bar{Y}}_{ST}) = \sum_{h=1}^{L} W_h^2 \cdot \left(\frac{\sum_{h=1}^{L} W_h \cdot S_h}{n \cdot W_h \cdot S_h} - N_h^{-1} \right) \cdot S_h^2$$

$$Var_o(\hat{\bar{Y}}_{ST}) = \sum_{h=1}^{L} \left(W_h^2 \cdot \frac{\sum_{h=1}^{L} W_h \cdot S_h}{n \cdot W_h \cdot S_h} \cdot S_h^2 - W_h^2 \cdot S_h^2 \cdot N_h^{-1} \right) =$$

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

Varianza dello stimatore della media con allocazione ottimale!!!

Per praticità, se scomponessimo anche la varianza dello stimatore della media con allocazione proporzionale:

$$Var_p\left(\hat{\bar{Y}}_{ST}\right) = \frac{1}{N^2} \cdot Var_p\left(\hat{Y}_{ST}\right) = \frac{1}{n}(1-f)\sum_{h=1}^{L} W_h \cdot S_h^2$$

$$Var_p\left(\hat{Y}_{ST}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{L} W_h \cdot S_h^2 = \frac{1}{n}\sum_{h=1}^{L} W_h \cdot S_h^2 - \frac{1}{N}\sum_{h=1}^{L} W_h \cdot S_h^2$$

Bisogna quantificare la relazione:

$$Var_p\left(\hat{\bar{Y}}_{ST}\right) - Var_o\left(\hat{\bar{Y}}_{ST}\right)$$

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

$$Var_p(\hat{\bar{Y}}_{ST}) - Var_o(\hat{\bar{Y}}_{ST}) = \frac{1}{n}\sum_{h=1}^{L} W_h \cdot S_h^2 - \frac{1}{n}\cdot\left(\sum_{h=1}^{L} W_h \cdot S_h\right)^2$$

$$Var_p(\hat{\bar{Y}}_{ST}) = \frac{1}{n}\sum_{h=1}^{L} W_h \cdot S_h^2 - \frac{1}{n}\cdot\left(\sum_{h=1}^{L} W_h \cdot S_h\right)^2 + Var_o(\hat{\bar{Y}}_{ST})$$

La varianza dello stimatore con allocazione ottimale (o allocazione di Neyman) è più piccola di quella che si consegue con l'allocazione proporzionale.

Il guadagno di precisione
è tanto più elevato quanto più variabile
si presenta la dispersione del carattere tra gli strati

Stimatore per quoziente

1 - Stimatore per quoziente separato:

$$\hat{Y}_{qs} = \sum_{h=1}^{L} \frac{\hat{Y}_h}{\hat{X}_h} X_h$$

dove

$$\left.\begin{aligned} \hat{Y}_h &= N_h \cdot \bar{y}_h \\ \hat{X}_h &= N_h \cdot \bar{x}_h \end{aligned}\right\}$$ *Stimatori corretti del totale*

$$Var(\hat{Y}_{qs}) = \sum_{h=1}^{L} N_h^2 \frac{1-f_h}{n_h} \left(S_{hy}^2 + \hat{R}_h^2 S_{hx}^2 - 2\hat{R} S_{hxy} \right)$$

---

2 - Stimatore per quoziente combinato:

$$\hat{Y}_{qc} = \frac{\hat{Y}_{ST}}{\hat{X}_{ST}} X$$

dove

$$\left.\begin{aligned} \hat{Y}_{ST} &= \sum_{h=1}^{L} N_h \cdot \bar{y}_h \\ \hat{X}_{ST} &= \sum_{h=1}^{L} N_h \cdot \bar{x}_h \end{aligned}\right\}$$ *Stimatori corretti del totale nel campionamento stratificato*

Non richiede i totali di strato ma il totale dell'intera popolazione secondo la variabile ausiliaria X!!!

La varianza dello stimatore, per misurare l'efficienza dello stimatore per quoziente combinato, è

$$Var(\hat{Y}_{qc}) = Var(\hat{Y}_{ST}) + \hat{R}_c^2 \cdot Var(\hat{X}_{ST}) - 2\hat{R}_c \cdot Cov(\hat{Y}_{ST}, \hat{X}_{ST})$$

dove

$$\hat{R}_c = \frac{\hat{Y}_{ST}}{\hat{X}_{ST}}$$

$$Cov(\hat{X}_{ST}, \hat{Y}_{ST}) = N^2 \frac{1-f}{n} \sum_{h=1}^{L} W_h S_{hxy}$$

$$Var(\hat{X}_{ST}) = N^2 \frac{1-f}{n} \sum_{h=1}^{L} W_h S_h^2$$

Stimatore R

Stimatore per regressione

1 - Stimatore per regressione separato:

$$\hat{Y}_{rs} = \sum_{h=1}^{L} \hat{Y}_h + \beta_h (X_h - \hat{X}_h)$$

dove

$$\left.\begin{aligned} \hat{Y}_h &= N_h \cdot \bar{y}_h \\ \hat{X}_h &= N_h \cdot \bar{x}_h \end{aligned}\right\}$$ *Stimatori corretti del totale*

$$Var(\hat{Y}_{rs}) = \sum_{h=1}^{L} N_h^2 \frac{1 - f_h}{n_h} \left(S_{hy}^2 + \hat{\beta}_h^2 S_{hx}^2 - 2\hat{\beta} S_{hxy} \right)$$

---

2 - Stimatore per regressione combinato:

$$\hat{Y}_{rc} = \hat{Y}_{ST} + \hat{\beta}_c (X - \hat{X}_{ST})$$

dove

$$\left.\begin{aligned} \hat{Y}_{ST} &= \sum_{h=1}^{L} N_h \cdot \bar{y}_h \\ \hat{X}_{ST} &= \sum_{h=1}^{L} N_h \cdot \bar{x}_h \end{aligned}\right\}$$ *Stimatori corretti del totale nel campionamento stratificato*

$$\hat{\beta}_c = \frac{Cov(\hat{X}_{ST}, \hat{Y}_{ST})}{Var(\hat{X}_{ST})}$$

$$Cov(\hat{X}_{ST}, \hat{Y}_{ST}) = N^2 \frac{1-f}{n} \sum_{h=1}^{L} W_h S_{hxy}$$

$$Var(\hat{X}_{ST}) = N^2 \frac{1-f}{n} \sum_{h=1}^{L} W_h S_h^2$$

Stimatore beta

La varianza dello stimatore, per misurare l'efficienza dello stimatore per regressione combinato, è

$$Var(\hat{Y}_{rc}) = Var(\hat{Y}_{ST}) + \hat{\beta}_c^2 \cdot Var(\hat{X}_{ST}) - 2\hat{\beta}_c \cdot Cov(\hat{Y}_{ST}, \hat{X}_{ST})$$

ESERCITAZIONE – CAMPIONAMENTO STRATIFICATO

Esercizio 1

Allo scopo di stimare l'andamento del mercato immobiliare, un'agenzia operante in Campania effettua un'indagine campionaria scegliendo casualmente 30 comuni della regione. Nella tabella viene riportata la ripartizione dei comuni della regione per ampiezza demografica e la spesa media per immobili e la relativa varianza calcolata sull'indagine campionaria che viene effettuata in ogni area.

	N_h	$\overline{y}_h$	S^2_h
<5000	*126*	*120*	*85*
5000-15.000	*92*	*63*	*76*
>15.000	*75*	*60*	*175*

Volendo effettuare un campionamento stratificato di numerosità n=30, si stabilisca:

a) Se sia opportuno procedere con un campionamento stratificato
b) come ripartire tale numerosità nei tre diversi strati secondo la ripartizione uniforme, proporzionale e quella ottimale (ipotizzando solo in questo caso che s_h siano informazioni della popolazione);
c) gli stimatori del totale e della media e le relative varianze nel caso di allocazione proporzionale;

Esercizio 2

La direzione generale di un istituto bancario a diffusione nazionale desidera valutare l'opinione dei propri dipendenti in merito all'apertura degli sportelli nella mattina del sabato; decide quindi di selezionare casualmente 32 agenzie delle 121 presenti sul territorio e ripartite in macrozone come riportato nella seguente tabella:

Macrozone	*Nord*	*Centro*	*Sud*	*Isole*
Totale agenzie	*60*	*34*	*19*	*8*

Nell'ipotesi di intervistare tutti i dipendenti delle 32 agenzie scelte, la rilevazione ha fornito la seguente proporzione di favorevoli all'apertura degli sportelli di sabato mattina:

Macrozone	*Nord*	*Centro*	*Sud*	*Isole*
Prop favorevoli	*0,12*	*0,11*	*0,08*	*0,07*

a) Si definisca il piano di campionamento e si effettui l'allocazione più opportuna per il campione.
b) Si stimi la proporzione di dipendenti favorevoli all'apertura straordinaria del sabato mattina.

Esercizio 3

Si vuole condurre un'indagine nel Nord Italia sull'ammontare mensile di multe date per eccesso di velocità. Vengono scelti 40 comuni, suddivisi in tre aree.
Considerate le informazioni contenute nella seguente tabella, allocare il campione nel modo più opportuno e stimare l'ammontare mensile delle multe confrontando le due strategie suggerite dalle informazioni date.

	Popolazione		**Campione**				
	Numero comuni	**Totale pop residente (migliaia)**	**Numero multe**	**Popolazione residente nei comuni estratti**	**Sx^2**	**Sy^2**	**Sxy**
Area 1	*91*	*250*	*15*	*60*	*61*	*44*	*40,21*
Area 2	*36*	*180*	*25*	*65*	*60*	*79*	*54,76*
Area 3	*20*	*200*	*43*	*90*	*105*	*296*	*140,35*

Esercizio 4

La Circoscrizione di Bagnoli conta, secondo i dati del Censimento 2001, 24.700 residenti, di cui 11.839 uomini e 12.861 donne.
Il Consiglio circoscrizionale vuole indagare sul gradimento dei cittadini rispetto al nuovo assetto dell'illuminazione stradale e, perlomeno in una prima fase, verificare se complessivamente questi risultino favorevoli o contrari alla nuova situazione.
Si decide di procedere con una indagine campionaria per stimare la proporzione di cittadini favorevoli, ricorrendo ad un campione stratificato di tipo proporzionale con variabile di stratificazione rappresentata dal genere (Maschio/Femmina).

1. Si determini la numerosità campionaria necessaria per avere un errore massimo della stima di due punti percentuali (in più o in meno) ad un livello di fiducia del 90%;
2. si ripartisca il campione determinando il numero di maschi e di femmine che dovranno essere intervistati.

Supponiamo poi che nel campione si siano dichiarati favorevoli alla nuova illuminazione 446 uomini e 638 donne.

3. Si determini la stima complessiva della proporzione di cittadini favorevoli e la relativa varianza