Campionamento a Grappoli:

- Campionamento a grappoli;
- Differenze rispetto al campionamento stratificato;
- Pregi e difetti;
- Fasi di realizzazione;
- Estrazione senza ripetizione a probabilità costanti (grappoli della stessa ampiezza);
- Estrazione senza ripetizione a probabilità variabili (grappoli di diversa ampiezza)
- Estrazione con ripetizione a probabilità variabili (grappoli di diversa ampiezza)
- Unità autorappresentative;
- Stratificazione dei grappoli;

Campionamento a grappoli:

A volte, le unità elementari di un collettivo non vengono selezionate una ad una (come accade nel campionamento casuale semplice), ma per gruppi.

Ciò avviene quando le unità statistiche della popolazione sono raggruppate in insiemi o aggregati, detti **grappoli**, per i quali si è di fronte a due problematiche procedurali:

a) **non esiste la lista** nominativa delle singole unità della popolazione da cui estrarre il campione e la formazione della stessa per la sola rilevazione in atto risulta molto onerosa;

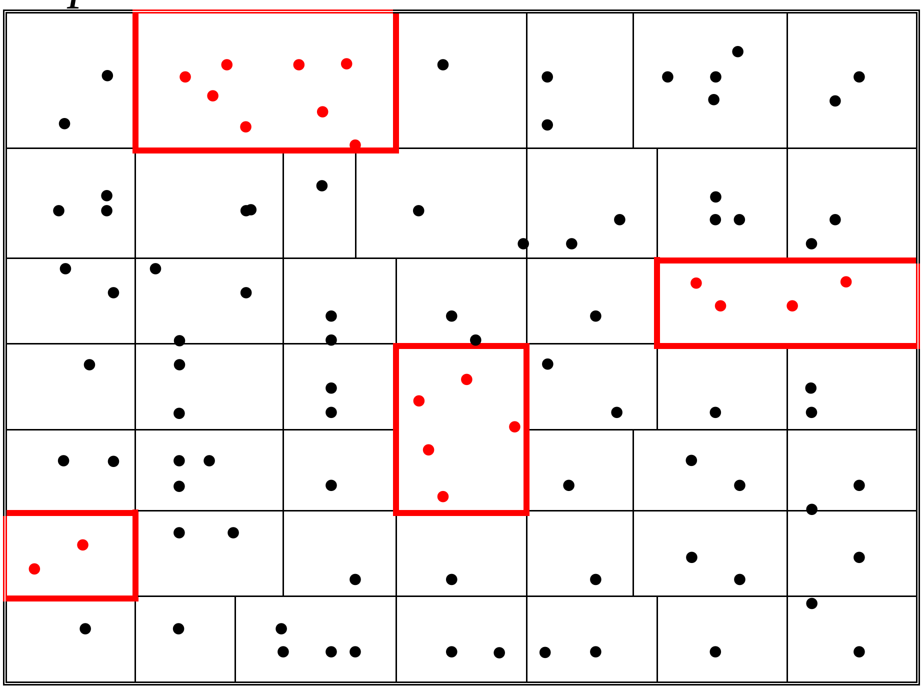
b) si ha interesse a **ridurre i costi** di rilevazione (l'osservazione di unità vicine è più economica e più facile di quella su unità sparse e distanti tra di loro).

La popolazione presente sul territorio italiano è la somma delle sottopopolazioni presenti sui territori regionali. All'interno di ciascuna regione, la popolazione è distribuita in province e, all'interno delle province, in comuni.

Gli studenti di un ateneo sono classificati in facoltà, quelli di una scuola, in classi, e così via dicendo.
Questi raggruppamenti di unità possono essere utilizzati come strati ma, alternativamente, possono essere utilizzati come unità di selezione e in questo caso sono denominati grappoli.
L'elenco dei grappoli forma la lista dalla quale viene estratto il campione.

Presupposto:
la popolazione è suddivisa, in modo naturale o artificiale, in sottoinsiemi di unità legati da vincoli di contiguità spaziale.

Popolazione



L'elenco dei grappoli rappresenta la lista dalla quale viene estratto il campione.

Ipotesi:

- *Il campione è costituito da TUTTE le unità incluse nel grappolo estratto (campionamento ad uno stadio);*

- *Il campione è costituito da ALCUNE unità estratte senza ripetizione all'interno del grappolo estratto (campionamento a due stadi).*

La differenza rispetto ai piani di campionamento precedenti è che in questo caso le unità vengono selezionate contemporaneamente (i gruppi) e non più una alla volta.

Le unità della popolazione sono i gruppi (insiemi di unità elementari) estratti senza o con ripetizione.

Caratteristica del piano di campionamento:

- *Eterogeneità all'interno dei grappoli;*
- *Omogeneità tra i grappoli;*
- *Ordinamento gerarchico delle unità.*

Il numero di stadi dipende da quello dei livelli gerarchici di aggregazione delle unità

Esempio (1): Indagine ISTAT sulle forze lavoro in Italia.

- Unità di *primo* stadio: **Comuni;**
- Unità di *secondo* stadio: **Famiglie;**
(selezionate dalle anagrafi dei Comuni selezionati)
- Unità di *terzo* stadio: **Componenti delle famiglie.**

Unità campionarie:
1. Famiglie ➔ *campionamento a due stadi;*
2. Componenti delle famiglie ➔ *campionamento a tre stadi;*

L'ISTAT SVOLGE UN PIANO DI CAMPIONAMENTO
A DUE STADI!!!

Esempio (2): Un'impresa commerciale con punti vendita al dettaglio sul territorio nazionale svolge un'indagine su un campione di propri clienti.

1. Creazione della lista delle unità costituenti la popolazione: cliente è colui che effettua almeno un acquisto presso un proprio punto vendita in un determinato periodo di riferimento.

- Unità di *primo* stadio: **Punti vendita;**
- Unità di *secondo* stadio: **Clienti;**

Differenze rispetto al campionamento stratificato

1. Nel campionamento a stadi i gruppi sono omogenei tra di loro ma eterogenei al loro interno. Il campionamento stratificato ha caratteristiche opposte.

2. Nel campionamento stratificato le unità vengono estratte da tutti i raggruppamenti (strati). Nel campionamento a stadi i grappoli vengono estratti esattamente come un campionamento casuale semplice.

   E' un CCS applicato ai gruppi che rappresentano la popolazione.

Rispetto al campionamento CCS, quello a stadi è più efficiente se la variabilità interna ai gruppi è più elevata rispetto a quella prevista dall'intera popolazione.

Nella realtà è meno efficiente di un CCS: nell'esempio delle famiglie o delle comunità, le unità elementari nella norma presentano una tendente omogeneità di comportamenti e di preferenze.

3. Nel campionamento a stadi la numerosità delle unità non è nota a priori. Si può stabilire il numero di grappoli da selezionare ma non le unità elementari da intervistare

Pregi:

1. *Costi bassi di rilevazione*: se i grappoli sono insiemi di unità territorialmente vicine, i costi e i tempi di spostamento da un'unità all'altra all'interno di un grappolo sono molto bassi
(una volta identificate l'unità di primo/secondo/terzo stadio, il costo di rilevazione e' all'incirca lo stesso);

2. *Poca conoscenza delle unità*: la rilevazione dei dati è concentrata soltanto sulle unità di primo stadio; richiede soltanto la lista dei grappoli e delle unità statistiche appartenenti ai soli grappoli estratti.

Difetti:
1. *Minore efficienza del campionamento stratificato*: raramente l'omogeneità fra i grappoli è perfetta;

2. *Lavoro preliminare*: può richiedere del lavoro preliminare alla rilevazione, soprattutto se i grappoli non sono precostituiti.

Fasi di realizzazione:

1. *Individuare il numero appropriato di stadi sui quali effettuare la selezione;*

2. *Identificare le caratteristiche per stratificare soprattutto le unità di primo stadio* (tale fase consente di migliorare le stime del piano di campionamento)*;*

3. *Decidere quante unità bisogna selezionare al primo e ai successivi stadi.*

4. Decidere come selezionare le unità.

Simbologia:
N – Numero di grappoli nella popolazione;
n – Numero di grappoli nel campione;

M_i – Numero di unità elementari nel grappolo i;

$M_.$ – Numero di unità elementari nella popolazione

$$M_\bullet = M_1 + M_2 + \dots\dots + M_N = \sum_{i=1}^{N} M_i$$

Y_{ij} – Valore della j-esima unità nel grappolo i della popolazione;

y_{ij} – Valore della j-esima unità nel grappolo i del campione;

$Y_{i.}$ – Totale della variabile Y nel grappolo i della popolazione;

$$Y_{i\bullet} = Y_{i1} + Y_{i2} + \dots\dots + Y_{iM} = \sum_{j=1}^{M_i} Y_{ij}$$

$y_{i.}$ – Totale della variabile Y nel grappolo i del campione;

$$y_{i\bullet} = y_{i1} + y_{i2} + \dots\dots + y_{iM} = \sum_{j=1}^{M_i} y_{ij}$$

Sintesi:

Grappoli	*Elementi*	*Numerosità*	*Totale*
1	$Y_{11}, Y_{12}, \ldots, Y_{1j}, \ldots, Y_{1M}$	M_1	$Y_{1.}$
2	$Y_{21}, Y_{22}, \ldots, Y_{2j}, \ldots, Y_{2M}$	M_2	$Y_{2.}$
:	:	:	
I	$Y_{i1}, Y_{i2}, \ldots, Y_{ij}, \ldots, Y_{iM}$	M_i	$Y_{i.}$
:	:	:	
N	$Y_{N1}, Y_{N2}, \ldots, Y_{Ni}, \ldots, Y_{NM}$	M_N	$Y_{N.}$

Totale della popolazione:

$$Y_{\bullet\bullet} = Y_{1\bullet} + Y_{2\bullet} + \ldots.. + Y_{N\bullet} = \sum_{i=1}^{N} Y_{i\bullet}$$

Estrazione senza ripetizione a probabilità costanti (grappoli della stessa ampiezza)

$$M_1 = M_2 = \ldots\ldots = M_N = M$$

I grappoli sono composti da un numero di unità tendenzialmente omogeneo!!!

- **stesse caratteristiche del CCS!!**

Le probabilità di inclusione sono costanti per tutte le unità della popolazione:

$$\pi = \frac{n}{N}$$

Stimatore corretto del totale:

$$\boxed{\hat{Y}_G = \sum_{i=1}^{n} \frac{y_{i\bullet}}{\pi_i} = \frac{N}{n} \sum_{i=1}^{n} y_{i\bullet} = N \cdot \bar{y}}$$

Stimatore corretto della media:

$$\boxed{\hat{\bar{Y}}_G = \frac{1}{N}\hat{Y}_G = \frac{1}{N}\cdot N\cdot\bar{y} = \bar{y}}$$

Media per grappolo

oppure

$$\boxed{\hat{\bar{\bar{Y}}}_{G(i)} = \frac{1}{M_{\bullet}}\cdot\hat{Y}_G = \frac{1}{N\cdot M}\cdot N\cdot\bar{y} = \frac{\bar{y}}{M}}$$

La relazione tra le due medie è:

$$\hat{Y}_G = N\cdot\hat{\bar{Y}}_G \quad ; \quad \hat{Y}_G = M_{\bullet}\cdot\hat{\bar{\bar{Y}}}_{G(i)} \Rightarrow N\cdot\hat{\bar{Y}}_G = M_{\bullet}\cdot\hat{\bar{\bar{Y}}}_{G(i)}$$

Dalla quale è possibile esprimere l'una in funzione dell'altra.

Ricordando che, per un campionamento casuale semplice a probabilità costanti:

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

la cui stima corretta è:

$$\boxed{s^2(\hat{Y}_G) = N^2 \frac{1-f}{n} \frac{\sum_{i=1}^{n}\left(y_{i\bullet} - \sum_{i=1}^{n} y_{i\bullet}/n\right)^2}{n-1}}$$

Stimatore corretto della varianza

dove:

f - *frazione di campionamento dei grappoli estratti*

$y_{i\bullet}$ - *totale del grappolo i-esimo della variabile Y*

Varianza dello stimatore della media:

1)

$$\hat{\bar{\bar{Y}}}_G = \frac{1}{N}\hat{Y}_G = \frac{1}{N}\cdot N\cdot\bar{y} = \bar{y}$$

Media per grappolo

$$\boxed{Var(\hat{\bar{\bar{Y}}}_G) = \frac{1}{N^2}Var(\hat{Y}_G) = \frac{1}{N^2}N^2\frac{1-f}{n}S_y^2 = \frac{1-f}{n}S_y^2}$$

2)

$$\hat{\bar{\bar{\bar{Y}}}}_{G(i)} = \frac{1}{M_\bullet}\cdot\hat{Y}_G = \frac{1}{N\cdot M}\cdot N\cdot\bar{y} = \frac{\bar{y}}{M}$$

$$\boxed{Var(\hat{\bar{\bar{\bar{Y}}}}) = \frac{1}{M_\bullet^2}Var(\hat{Y}_G) = \frac{1}{N^2\cdot M^2}Var(\hat{Y}_G) = \frac{1-f}{M^2\cdot n}S_y^2}$$

La varianza S^2 può essere espressa anche in termini di variabilità TRA i grappoli nel modo che segue.

$$S_y^2 = \sigma_1^2 = \frac{1}{N-1}\sum_{j=1}^{N}\left(Y_{j\bullet} - \overline{Y}_G\right)^2 =$$

$$= \frac{1}{N-1}\sum_{j=1}^{N}\left(M\cdot\overline{Y}_{j\bullet} - M\cdot\overline{\overline{Y}}\right)^2 = \frac{M^2}{N-1}\underbrace{\sum_{j=1}^{N}\left(\overline{Y}_{j\bullet} - \overline{\overline{Y}}\right)^2}_{S^2_B}$$

$$S_y^2 = \sigma_1^2 = M^2 \cdot S_B^2$$

da cui:

$$\boxed{Var(\hat{\overline{\overline{Y}}}) = \frac{1}{M_{\bullet}^2}Var(\hat{Y}_G) = \frac{1}{M^2}\cdot\frac{1-f}{n}\cdot\frac{M^2}{N-1}\sum_{j=1}^{N}\left(\overline{Y}_{j\bullet} - \overline{\overline{Y}}\right)^2 = \frac{1-f}{n}S_B^2}$$

Una stima corretta delle funzione precedente puo' essere:

$$s_y^2 = \hat{\sigma}_1^2 = M^2\cdot s_B^2 = M^2\cdot\frac{1}{n-1}\sum_{i=1}^{n}\left(\overline{y}_{j\bullet} - \overline{\overline{y}}\right)^2$$

Coefficiente di omogeneità nei grappoli:

Per valutare l'effetto del disegno di campionamento rispetto al campionamento casuale semplice, è necessario esprimere la varianza totale in termini di ***coefficiente di omogeneità nei grappoli***.

$$(NM-1)\cdot S_T^2 = \sum_{j=1}^{N}\sum_{i=1}^{M}\left(Y_{ij}-\overline{Y}_{G(i)}\right)^2$$

$$N(M-1)\cdot S_W^2 = \sum_{j=1}^{N}\sum_{i=1}^{M}\left(Y_{ij}-\overline{Y}_i\right)^2$$

$$(N-1)\cdot S_B^2 = M^2\sum_{j=1}^{N}\left(\overline{Y}_j-\overline{Y}_{G(i)}\right)^2$$

Dove la devianza totale è la somma della devianza tra i grappoli e la devianza entro i grappoli

$$(NM-1)\cdot S_T^2 = N(M-1)\cdot S_W^2 + (N-1)\cdot S_B^2$$

Da cui il coefficiente di omogeneità nei grappoli è dato dalla relazione

$$\delta = \frac{S_B^2}{S_T^2} = 1-\frac{S_W^2}{S_T^2}$$

Tale coefficiente è facilmente interpretabile in quanto compreso tra 0 e 1 ma potendo essere calcolato anche in

altra espressione, l'estremo inferiore può essere espresso in forma differente
Essendo:

$$(N-1)\cdot S_B^2 \geq 0$$

Ne consegue che:

$$\frac{S_W^2}{S_T^2} \leq \frac{(NM-1)}{N(M-1)}$$

$$\delta = 1 - \frac{S_W^2}{S_T^2} \geq 1 - \frac{(NM-1)}{N(M-1)} = -\frac{N-1}{N(M-1)}$$

In definitiva

$$-\frac{N-1}{N(M-1)} \leq \delta \leq 1$$

o anche in forma approssimata

$$-\frac{1}{M-1} \leq \delta \leq 1$$

Se **$S^2_W=0$:** massima omogeneità tra le unità nei grappoli *(la variabilità è dovuta soltanto alla varianza tra i grappoli).*
Se i grappoli sono omogenei, vale a dire formati da unità elementari molti simili tra loro per la variabile di interesse, il valore di δ tende all'estremo superiore .

Questo caso è rappresentato nella tabella seguente dove le 36 unità di una ipotetica popolazione sono ripartite in 6 grappoli di uguale dimensione.

GR1	GR2	GR3	GR4	GR5	GR6
150	*160*	*170*	*180*	*190*	*200*
150	*160*	*170*	*180*	*190*	*200*
150	*160*	*170*	*180*	*190*	*200*
150	*160*	*170*	*180*	*190*	*200*
150	*160*	*170*	*180*	*190*	*200*
150	*160*	*170*	*180*	*190*	*200*

Poiché i grappoli sono composti da unità elementari identiche non si possono considerare rappresentativi della popolazione; infatti l'estrazione di uno di essi equivale all'estrazione di una unità.
Questa è la situazione peggiore del campionamento.

Se **S^2_B=0:** massima eterogeneità tra le unità nei grappoli *(la variabilità è dovuta soltanto alla varianza all'interno dei grappoli).*

Se i grappoli sono altamente eterogenei al loro interno, il valore di δ tende all'estremo inferiore della disuguaglianza. Questo caso viene rappresentato nella tabella che segue:

GR1	GR2	GR3	GR4	GR5	GR6
150	*150*	*150*	*150*	*150*	*150*
160	*160*	*160*	*160*	*160*	*160*
170	*170*	*170*	*170*	*170*	*170*
180	*180*	*180*	*180*	*180*	*180*
190	*190*	*190*	*190*	*190*	*190*
200	*200*	*200*	*200*	*200*	*200*

I grappoli solo perfettamente identici e rappresentativi della popolazione ed è sufficiente l'estrazione di uno di essi perché il campionamento sia più efficiente di un campionamento casuale semplice.

Stima delle varianze:

Una stima corretta delle funzioni di varianza NEI gruppi e TRA i gruppi e':

$$s_T^2 = \frac{N(M-1)\cdot s_W^2 + M(N-1)\cdot s_B^2}{(NM-1)}$$

$$s_B^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{y}_{j\bullet} - \bar{y}\right)^2$$

$$s_W^2 = \frac{1}{n(M-1)} \sum_{i=1}^{n} \sum_{j=1}^{M} \left(y_{ij} - \bar{y}_i \right)^2$$

Stimatore corretto della proporzione:

$$\boxed{\hat{P}_{GR} = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_i}$$

Media delle proporzioni dei grappoli

dove

$$\hat{p}_i = \frac{a_i}{M}$$

La varianza dello stimatore risulta:

$$Var(P_{GR}) = \frac{1-f}{n} \frac{\sum_{i=1}^{n} \left(\hat{p}_i - \hat{P}_{GR} \right)^2}{n-1}$$

Estrazione senza ripetizione a probabilità variabili (grappoli di diversa ampiezza)(1)

$$M_1 \neq M_2 \neq \ldots\ldots \neq M_N$$

Hp: *Le ampiezze dei grappoli non risultano note a priori:*

La selezione dei grappoli avviene con tecnica casuale semplice senza ripetizione a probabilità costanti!!

$$\pi = \frac{n}{N}$$

Stimatore corretto del totale:

$$\boxed{\hat{Y}_G = \sum_{i=1}^{n} \frac{y_{i\bullet}}{\pi_i} = \frac{N}{n} \sum_{i=1}^{n} \sum_{k=1}^{M_i} y_{ik} = \frac{N}{n} \sum_{i=1}^{n} y_{i\bullet} = N \cdot \bar{y}}$$

Ricordando che, per un campionamento casuale semplice a probabilità costanti:

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

Stimatore corretto della media:

Problemi nella stima della media per unità elementare:

$$\boxed{\hat{\overline{Y}}_G = \frac{1}{N}\hat{Y}_G = \frac{1}{N} \cdot N \cdot \overline{y} = \overline{y}}$$

Media per grappolo

oppure

$$\boxed{\hat{\overline{\overline{Y}}} = \frac{1}{M_{\bullet}} \cdot \hat{Y}_G = \frac{\hat{Y}_G}{\sum_{j=1}^{N} M_j} = \frac{\hat{Y}_G}{N \cdot \overline{\overline{M}}} = \frac{\overline{y}}{\overline{\overline{M}}}}$$

Media per unità elementari

Ricordando che siamo nell'ipotesi di non conoscenza delle ampiezze dei grappoli, una stima alternativa puo' essere fornita dallo stimatore per quoziente della media.

Stimatore per quoziente:

Se i totali dei grappoli sono correlati positivamente con le ampiezze M_i allora lo stimatore per quoziente è più efficiente dello stimatore corretto:

Ricordando che:

$$\hat{Y}_q = \frac{\hat{Y}}{\hat{X}} X \quad e \quad \hat{Y} = \frac{N}{n} \sum_{i=1}^{n} y_{i\bullet}$$

$$\boxed{\hat{Y}_{Gq} = \frac{\sum_{i=1}^{n} y_{i\bullet}}{\sum_{i=1}^{n} M_i} M_\bullet}$$

Stimatore per quoziente del totale.

Inoltre:

$$Var(\hat{Y}_{Gq}) = N^2 \frac{1-f}{n} \frac{\sum_{j=1}^{N} \left(Y_{j\bullet} - R \cdot X_j\right)^2}{N-1}$$

se:

$$X_i = M_i \quad ; \quad Y_i = Y_{i\bullet} \quad ; \quad R = \frac{Y_{\bullet\bullet}}{M_\bullet} = \bar{Y}_{\bullet\bullet}$$

$$\boxed{Var(\hat{Y}_{Gq}) = N^2 \frac{1-f}{n} \frac{\sum_{j=1}^{N} \left(Y_{j\bullet} - M_j \cdot \bar{Y}_{\bullet\bullet}\right)^2}{N-1}}$$

Varianza dello stimatore

Ricordando che:

$$x_i = M_i \quad ; \quad \bar{y} = \frac{\sum_{i=1}^{n} y_{i\bullet}}{\sum_{i=1}^{n} M_i} \quad ; \quad R = \frac{\sum_{i=1}^{n} y_{i\bullet}}{\sum_{i=1}^{n} M_i}$$

$$\boxed{s^2(\hat{Y}_{Gq}) = N^2 \frac{1-f}{n} \frac{\sum_{i=1}^{n} \left(y_{i\bullet} - M_i \cdot \bar{y}\right)^2}{n-1}}$$

Stimatore corretto della varianza

Che puo' essere scritto nella sua forma alternativa:

$$s^2(\hat{Y}_{Gq}) = N^2 \frac{1-f}{n} \frac{\sum_{i=1}^{n} M_i^2 (\bar{y}_{i\bullet} - \bar{y})^2}{n-1}$$

Lo stimatore medio per quoziente, per ogni unità elementare, è invece

$$\boxed{\hat{\bar{Y}}_{Gq} = \frac{\hat{Y}_{Gq}}{M_\bullet} = \frac{\sum_{i=1}^{n} y_{i\bullet}}{\sum_{i=1}^{n} M_i}}$$

Stimatore per quoziente della media.

Non dipende dal totale della popolazione ma soltanto dall'ampiezza delle unita' campionate .

PROBLEMA: varianza dello stimatore e' in funzione di M•!!!

$$\boxed{s^2(\hat{\bar{Y}}_{Gq}) = \frac{s^2(\hat{Y}_{Gq})}{M_\bullet^2} = \frac{N^2}{M_\bullet^2} \frac{1-f}{n} \frac{\sum_{i=1}^{n} (y_{i\bullet} - M_i \cdot \bar{y})^2}{n-1}}$$

Varianza dello stimatore per quoziente della media.

Il totale della dimensione M della popolazione si sostituisce con una sua stima corretta

$$M_{\bullet} = \sum_{i=1}^{n} M_i$$

Stimatore corretto della proporzione:
Se i grappoli hanno dimensioni diverse, lo stimatore della proporzione diventa:

$$\boxed{\hat{P}_{GR} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} M_i}}$$

Dove la proporzione dell'i-esimo grappolo è:

$$\hat{p}_i = \frac{a_i}{M_i}$$

La relativa varianza dello stimatore risulta:

$$Var(P_{GR}) = \frac{1-f}{n\overline{M}^2} \frac{\sum_{i=1}^{n} \left(a_i - \hat{P}_{GR} \cdot M_i\right)^2}{n-1}$$

Estrazione con ripetizione a probabilità variabili (grappoli di diversa ampiezza) (2)

$$M_1 \neq M_2 \neq \ldots\ldots \neq M_N$$

Hp: *Le ampiezze dei grappoli risultano note a priori:*

Stimatore corretto:

$$\boxed{\hat{Y}_{PPS} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_{i\bullet}}{p_i}}$$

Stimatore per espansione del totale.

se $P_i = M_i / M_\bullet$

$$\boxed{\hat{Y}_{PPS} = M_\bullet\left(\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i\bullet}\right)}$$

Ricordando che, per un campionamento a probabilità variabili con ripetizione:

$$Var(\hat{Y}) = \frac{1}{n}\sum_{j=1}^{N}\left(\frac{Y_j}{P_j} - Y\right)^2 \cdot P_j$$

Nel campionamento a grappoli diventa:

$$Var(\hat{Y}_{PPS}) = \frac{1}{n}\sum_{j=1}^{N}\left(\frac{Y_{j\bullet}}{P_j} - Y_{\bullet\bullet}\right)^2 \cdot P_j$$

Varianza dello stimatore totale

$$se \quad P_i = M_i / M_\bullet$$

$$Var(\hat{Y}_{PPS}) = \frac{1}{n}\sum_{j=1}^{N}\left(\frac{M_\bullet Y_{j\bullet}}{M_j} - M_\bullet \bar{\bar{Y}}_{\bullet\bullet}\right)^2 \cdot \frac{M_j}{M_\bullet} =$$

$$Var(\hat{Y}_{PPS}) = \frac{1}{n}\sum_{j=1}^{N}\frac{M_j}{M_\bullet} M_\bullet^2 \left(\bar{Y}_{j\bullet} - \bar{\bar{Y}}_{\bullet\bullet}\right)^2 =$$

$$\boxed{Var(\hat{Y}_{PPS}) = \frac{M_\bullet}{n}\sum_{j=1}^{N} M_j \cdot \left(\bar{Y}_{j\bullet} - \bar{\bar{Y}}_{\bullet\bullet}\right)^2}$$

Stima della varianza:

Ricordando che, per un campionamento a probabilità variabili con ripetizione:

$$s^2(\hat{Y}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{Y}\right)^2$$

Nel campionamento a grappoli diventa:

$$s^2(\hat{Y}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_{i\bullet}}{p_i} - \hat{Y}_{PPS}\right)^2$$

$$se \quad P_i = M_i / M_\bullet$$

$$s^2(\hat{Y}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{M_\bullet \cdot y_{i\bullet}}{M_i} - M_\bullet \cdot \hat{\bar{\bar{Y}}}_{PPS}\right)^2$$

$$s^2(\hat{Y}_{PPS}) = \frac{1}{n(n-1)}\sum_{i=1}^{n} M_\bullet^2\left(\bar{y}_{i\bullet} - \left(\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i\bullet}\right)\right)^2$$

$$\boxed{s^2(\hat{Y}_{PPS}) = \frac{M_\bullet^2}{n(n-1)}\sum_{i=1}^{n}\left(\bar{y}_{i\bullet} - \frac{\hat{Y}_{PPS}}{M_\bullet}\right)^2}$$

Estrazione senza ripetizione a probabilità variabili (grappoli di diversa ampiezza) (2)

$$M_1 \neq M_2 \neq \ldots\ldots \neq M_N$$

Hp: *Le ampiezze dei grappoli risultano note a priori:*

Stimatore corretto:

$$\boxed{\hat{Y}_{PPS} = \sum_{i=1}^{n} \frac{y_{i\bullet}}{\pi_i}}$$

Stimatore per espansione del totale.

dove le probabilità di inclusione di primo ordine vanno calcolate con le tecniche note di estrazione a probabilità variabili.

$$se \quad P_i = M_i / M_\bullet$$

e la varianza dello stimatore risulta:

$$\boxed{s^2(\hat{Y}_{PPS}) = \sum_{i=1}^{n} \frac{1-\pi_i}{\pi_i^2} \cdot y_{i\bullet} + 2\sum_{i=1}^{n} \sum_{k>i}^{n} \frac{\pi_{ik} - \pi_i \cdot \pi_k}{\pi_{ik} \cdot \pi_i \cdot \pi_k} \cdot y_{i\bullet} \cdot y_{k\bullet}}$$

Disegni con stratificazione delle unità primarie:

Ogni campione ad uno o a più stadi è generalmente stratificato al primo stadio!!!

Logica di stratificazione

Logica della determinazione ottimale dei confini (MAHALANOBIS): delimitare gli strati in modo che la dimensione sia all'incirca costante.

$$M_h = \frac{M_\bullet}{L}$$

Numero ideale di unita' di secondo stadio da includere in ogni strato!!

Stratificare le unità di primo stadio vuol dire aumentare l'efficienza delle stime del disegno di campionamento adottato.

Unità autorappresentative:

Determinazione delle unità primarie che possono essere collocate in uno strato a se stante e selezionate con probabilità 1.

Valore soglia di autorappresentatività

$$M^* = \frac{M_\bullet}{n}$$

Le unità primarie la cui dimensione è maggiore di tale valore soglia possono essere giudicate autorappresentative.

Vantaggi:

1. aumento dell'efficienza attesa delle stime;
2. migliore organizzazione del lavoro sul campo.

Stratificazione dei grappoli

Non esiste disegno campionario che non preveda la stratificazione delle unita' primarie (grappoli)

MIGLIORAMENTO DELLE STIME!!!

↓

Misurazione del guadagno di efficienza

↓

Hp1: ***Gli N grappoli sono della stessa ampiezza;***
Hp2: ***Allocazione proporzionale dei grappoli negli strati.***

Stimatori del totale

$$\hat{Y}_{ST} = \sum_{h=1}^{L} \hat{Y}_h$$

con

Y_h- stimatore del totale dello strato h-esimo

Le varianze degli stimatori del totale sono rispettivamente

$$Var(\hat{Y}_{ST}) = \sum_{h=1}^{L} Var_h(\hat{Y}_h)$$

e quella corretta:

$$s^2(\hat{Y}_{ST}) = \sum_{h=1}^{L} s_h^2(\hat{Y}_h)$$

Guadagno di efficienza:

Varianza dello stimatore del totale nel campionamento stratificato

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

dove

$$S_h^2=\frac{\sum_{j=1}^{N_h}\left(Y_{hj\bullet}-Y_{h\bullet\bullet}/N_h\right)^2}{N_h-1}$$

con

$Y_{hj.}$- totale del grappolo j-esimo nello strato h;

$Y_{h..}$- totale di strato.

Varianza dello stimatore del totale nel campionamento a grappoli (senza strati)

$$Var\left(\hat{Y}_G\right)=\frac{N^2}{n}(1-f)\frac{\sum_{h=1}^{L}\sum_{j=1}^{N_h}\left(Y_{hj\bullet}-Y_{\bullet\bullet\bullet}/N\right)^2}{N-1}$$

con

$Y_{...}$- totale della popolazione.

Esercizi - Campionamento a GRAPPOLI -

Esercizio 1:
Si consideri una popolazione di N=16 Comuni ciascuno composto da M=16 unità elementari.

i	yi	σ_i^2
1	2	0,117
2	19	5,496
3	27	7,429
4	19	1,763
5	23	1,729
6	40	6,933
7	43	11,300
8	40	4,000
9	39	3,196
10	64	7,600
11	44	8,333
12	35	4,963
13	34	4,783
14	22	1,717
15	38	9,717
16	25	4,263

Si supponga di aver estratto i seguenti grappoli:

1-7-10-15.

Calcolare lo stimatore della media per grappolo e per unità elementare e le relative varianze.

Esercizio 2:
Sia data una popolazione di N=10 aziende formate ciascuna da 3 reparti, per i quali si rileva il consumo in Kilowatt, se ne estraggono 4 come di seguito riportate (in parentesi le varianze campionarie dei gruppi).

Azienda 1	*Azienda 2*	*Azienda 3*	*Azienda 4*
1,4	1,5	1,25	1,9
1,2	1,6	1,85	2,1
1,3	1,8	1,65	1,8
0.01	*0.023*	*0.093*	*0.023*

a) Si calcoli il coefficiente di omogeneità nei grappoli dopo aver calcolato la varianza tra i gruppi, nei gruppi e quella totale.
b) Ipotizzando di estrarre un campione di eguale dimensione con tecnica casuale semplice (le prime di ogni azienda), effettuare un confronto di efficienza tra i due piani di campionamento.

Esercizio 3:
In un'indagine campionaria si vuole stimare il totale dei ricoveri giornalieri dei pazienti affetti da una determinata patologia nell'arco di un mese. Si estrae un campione SCR pari a n=4 sui 15 istituti di cura generali di cui si conoscono i rispettivi posti letto:

Istituto	*N.posti letto*	*N.Ricoveri*
1	650	55
2	700	48
3	290	16
4	360	21

Stimare il totale dei ricoveri nel piano di campionamento a grappoli con ripetizione sapendo che il numero di posti letto totali per i 15 istituti ospedalieri è di 5735.

Esercizio 4:
Si vuole condurre un'indagine campionaria sulle ore di assenza dal lavoro da parte degli addetti di 430 imprese di un particolare settore della regione Lombardia. Si supponga che vengano estratte senza ripetizione 5 imprese e si considerano tutti gli addetti di ciascuna impresa giungendo ai dati esposti in tabella:

Imprese	*N.addetti*	*Ore assenza*
1	17	45
2	15	30
3	29	49
4	12	20
5	34	26

a) Definire il piano di campionamento utilizzato;
b) Stimare il totale di ore di assenza degli addetti nelle 430 aziende con lo stimatore corretto e la corrispondente varianza;
c) Si dica se i dati a disposizione consentirebbero la stima del totale con lo stimatore per quoziente;
d) Supposto che il totale degli addetti nelle aziende in esame sia di 24515, calcolare la stima per quoziente e la relativa varianza.

Esercizio 5:
Ad una Asl afferiscono 15 poliambulatori con un totale di 227 dipendenti tra personale medico e paramedico. Per conoscere il numero totale di prestazioni mutualistiche effettuate in una settimana tipo si estrae un campione SSR a probabilità costanti di 3 poliambulatori rilevando i seguenti dati:

Poliambulatori	*N.dipendenti*	*N.prestazioni mutualistiche*
1	17	45
2	15	30
3	22	73

a) Stimare il numero totale di prestazioni mutualistiche nella settimana di riferimento per tutti i poliambulatori con la strategia che si ritiene più efficiente.

Esercizio 6:
Per conoscere con un certo anticipo il risultato delle ultime elezioni comunali, sono state scelte tre sezioni elettorali che hanno fornito, in tempi molto brevi, i risultati riportati di seguito:

Lista	*Sezione 1 % voti*	*Sezione 2 % voti*	*Sezione 3 % voti*
A	15	27	21
B	18	15	16
C	23	18	28
D	24	40	35

Sapendo che il comune è ripartito in 25 sezioni e che mediamente una sezione è composta da 500 elettori, si effettui una stima della proporzione in base ai voti rilevati da ciascuna lista.

Esercizio 7
Un'azienda che produce abbigliamento da lavoro fornisce tute agli operai di 14 stabilimenti industriali il cui numero di dipendenti è distribuito come riportato nella tabella.

Stab.	**Dip**
1	25
2	33
3	25
4	102
5	38
6	45
7	60
8	108
9	93
10	83
11	47
12	93
13	124
14	80

L'azienda ha proposto due modelli di tute, indicati con A e B, date in prova per una settimana. Allo scopo di mettere in produzione il modello più gradito, l'azienda intervista gli operai di 3 stabilimenti scelti casualmente.
Vengono estratti gli stabilimenti 2, 8 e 14 e si rileva il numero di operai che hanno dichiarato di preferire il modello di tuta A

Stabilimento	*Mi*	*ai*
2	33	28
8	108	75
14	80	53

Sulla base dei dati campionari, si stimi la proporzione relativa a tutta l'azienda e si determini l'errore standard.

Esercizio 8
In una indagine sulle presenze nelle strutture alberghiere di una città turistica durante una settimana di alta stagione, vengono estratti, con un campione SSR, 4 alberghi tra i 60 presenti sul territorio; i risultati ottenuti sono i seguenti:

Alberghi	*A*	*B*	*C*	*D*
n.° totale ospiti	380	420	125	84
n.°tot. pernottamenti	1634	2079	485	420

Sapendo che il numero totale dei posti letto è 14.800:
a. Stimare, con le strategie a disposizione il numero medio di giornate di soggiorno per ospite, nella settimana di riferimento e scegliere quella definitiva;
b. Indicare le caratteristiche delle strategie adottate;