

# Piani di campionamento a probabilità variabili

Obiettivo principale: ottenere, attraverso l'impiego di diversi metodi di estrazione e di selezione delle unità, **campioni più efficienti a parità di costo o campioni che a parità di efficienza, presentino costi più ridotti**

La caratteristica dei piani di campionamento a probabilità variabile è che alle unità della popolazione vengono associate probabilità di estrazione diverse

L'impiego delle probabilità variabili nel campionamento statistico fu introdotto da Hansen e Hurwitz nel 1943 proponendo uno schema di campionamento a due stadi con l'estrazione con ripetizione nel primo stadio

Il successo della proposta di Hansen e Hurwitz sta nel fatto che l'impiego delle probabilità variabili genera frequentemente **stimatori del totale** più efficienti rispetto all'uso delle probabilità costanti


# Piani di campionamento a probabilità variabili

Alcuni anni dopo la proposta di Hansen e Hurwitz, nel 1949, Mandow introdusse le probabilità variabili nel campionamento sistematico

Da allora furono proposti numerosi schemi di estrazione senza ripetizione con probabilità variabili a cominciare dallo schema di Narain introdotto nel 1951. Horvitz e Thompson nel 1952 formularono

**una teoria generale, nel caso di estrazioni con probabilità variabili**, basata sull'impiego di uno stimatore non distorto che prende il loro nome (HT):

- il principio è che per poter attribuire una probabilità variabile è necessario disporre di un indicatore di importanza relativa detto *misura di ampiezza o misura di dimensione*



# L'operazione di scelta delle unità della popolazione può essere di:

## Tipo probabilistico

### 1. Probabilità costanti *con e senza ripetizione*

- Semplice
- Stratificato
- A grappoli
- Sistemático
- A due o più stadi

### 2. Probabilità variabili *con o senza ripetizione*

- A grappoli\*
- Sistemático\*
- A due o più stadi\*
- Con probabilità di estrazione legata ad una variabile ausiliaria di dimensione



# Campionamenti probabilistici con probabilità variabili

Il campionamento casuale semplice ipotizza equiprobabilità delle unità estraibili, dei campioni e costanza delle probabilità di inclusione. Nel campionamento a probabilità variabile questa ipotesi è abbandonata.

## Esempio di utilizzo delle probabilità variabili

Indagine svolta nel 1989 sulle pari opportunità nell'industria metalmeccanica IT

La scelta delle  $n$  donne da intervistare (unità secondarie) veniva effettuata nelle unità locali scelte (primarie) fra tutte le  $N$  distribuite sul territorio nazionale.

# Campionamenti probabilistici con probabilità variabili

Poiché è noto che:

più è elevato il numero degli addetti



più è alta la frequenza di addetti di sesso femminile

**è evidente che per disporre di una buona presenza del genere femminile, alle unità locali con più addetti doveva essere assegnata una probabilità di estrazione maggiore**

# Criteri di classificazione

Attualmente si dispone di oltre sessanta schemi o metodi di campionamento con probabilità variabili che vengono classificati seguendo diversi criteri

Tra i criteri di classificazione più noti:

- Tipo di estrazione

- (i diversi metodi vengono raggruppati in funzione del tipo di estrazione)

- Classe equivalente riferito alle probabilità

- (si ipotizza l'identità delle probabilità di inclusione di tutte le possibili combinazioni di unità campionarie)

- Tipo di stimatore

- (diversi metodi vengono raggruppati in funzione dell'impiego di alcuni particolari e specifici stimatori)



# Criteri sul tipo di estrazione

## i. Metodi ad estrazione indipendente

(le unità della popolazione sono scelte una alla volta e ognuna con predefinita probabilità per ogni estrazione)

## ii. Metodi di rifiuto del campione

(prevedono un campionamento casuale con ripetizione con predefinite probabilità di estrazione per ogni unità, ma l'intero campione viene rifiutato ogniqualevolta si presenta il caso di una unità ripetuta)

## iii. Metodo di raggruppamenti in sottocampioni

(la popolazione viene suddivisa in sottopopolazioni in base a particolari criteri e viene estratto un elemento da ognuna di esse)

## iv. Altri metodi

(ad es. casi di estrazione dell'intero campione secondo una prefissata probabilità per ciascuno. Tale estrazione è progettata per tutti i campioni appartenenti all'intero spazio campionario)

# Notazioni

Sia  $X$  una variabile ausiliare ritenuta in relazione di approssimata proporzionalità con la variabile oggetto di studio  $Y$  (*nell'esempio l'indicatore di importanza relativa era dato dal numero di addetti nelle unità locali*). Il valore  $X_i$ , assunto da  $X$  nell'unità  $i$  della popolazione verrà chiamato *misura di ampiezza* (o di dimensione);

Il rapporto  $P_i = \frac{X_i}{X}, i = 1, 2, \dots, N$  dove  $X = \sum_1^N X_i$

è chiamato *misura di ampiezza normalizzata*

**Definizione**: Il campionamento con probabilità variabili, detto anche campionamento con probabilità proporzionali alla dimensione (PPS), consiste nell'estrarre, con ripetizione o senza ripetizione,  $n$  unità con probabilità proporzionali a una data misura d'ampiezza.

Tali misure di ampiezza costituiscono i *pesi* in base ai quali vengono selezionate le unità della popolazione.



# Tecniche di estrazione di singole unità campionarie $n=1$

## ■ Tecnica dei totali cumulati

è la tecnica operativa più comune per estrarre le unità di una popolazione con probabilità proporzionali alle misure di ampiezza  $X_i$

a) *Valori di ampiezza intera*

b) *Valori di ampiezza non intera*

## ■ Tecnica di Lahiri

ha lo stesso obiettivo ma risolve il problema della difficoltà della tecnica dei totali cumulati per  $N$  grande. Non è più necessario il calcolo delle misure di ampiezza cumulate, ma permette ugualmente di calcolare le probabilità di estrazione delle unità della popolazione proporzionali alle misure di ampiezza.

# La tecnica dei totali cumulati

a) *Valori di ampiezza intera*: Si calcolano le misure di ampiezza cumulate:

$$T_i = \sum_{j=1}^i X_j \quad i = 1, 2, \dots, N$$

$$X = \sum_{i=1}^N X_i$$

Si estrae quindi, per mezzo delle tavole dei numeri casuali, un numero casuale  $U$  con distribuzione uniforme discreta tra 1 e  $X = T_N$

L'unità scelta è la  $i$ -esima, se  $T_i$  è la prima  $T$  maggiore o uguale a  $U$ :  $T_{i-1} < U \leq T_i$

La probabilità di estrarre l'unità  $i$ -esima è:  $P_i = \frac{(T_i - T_{i-1})}{X} = \frac{X_i}{X}$

b) *Valori di ampiezza non intera*: Si calcolano le misure di ampiezza cumulate normalizzate:

$$Q_i = \sum_{j=1}^i \frac{X_j}{X} = \sum_{j=1}^i P_j \quad i = 1, 2, \dots, N$$

In tal caso il numero casuale  $U$  presenta una distribuzione rettangolare su  $[0,1]$  e se

$$Q_{i-1} < U \leq Q_i$$

la scelta cade sull'  $i$ -esima unità.

# La tecnica dei totali cumulati

a) Si calcolano le misure di ampiezza cumulate:

$$T_i = \sum_{j=i}^i X_j \quad i = 1, 2, \dots, N$$

Si estrae un numero casuale  $U$  con distribuzione uniforme discreta tra

$$1 \text{ e } X = T_N$$

L'unità scelta è la  $i$ -esima, se  $T_{i-1}$  è la prima  $T$  maggiore o uguale a  $U$ :

$$T_{i-1} < U \leq T_i$$

La probabilità di estrarre l'unità  $i$ -esima è :

$$P_i = \frac{(T_i - T_{i-1})}{X} = \frac{X_i}{X}$$

b) Si calcolano le misure di ampiezza cumulate normalizzate:

$$Q_i = \sum_{j=1}^i \frac{X_j}{X} = \sum_{j=1}^i P_j \quad i = 1, 2, \dots, N$$

In tal caso il numero casuale  $U$  presenta una distribuzione rettangolare su  $[0,1]$  e se

la scelta cade sull'  $i$ -esima unità.

$$Q_{i-1} < U \leq Q_i$$

Esempio:

Famiglie di uno stabile  $N=6$

Variabile analizzata  $Y$  = reddito (migliaia €)

Variabile ausiliare  $X$  = m<sup>2</sup> abitazione nota a priori da informazioni catastali

Unità	1	2	3	4	5	6
Variabili						
$Y_i$	1.8	2.1	2.4	3.0	3.5	3.7
$X_i$	45	65	60	80	85	120
$T_i$	45	110	170	250	335	455

Poiché le  $X_i$  sono intere considero il punto a)

Il numero casuale  $U$  da estrarre deve essere compreso nell'intervallo di numeri interi  $[1;X] = [1;455]$

Se si ottiene  $U = 147$ , l'unità scelta si identifica con la terza famiglia con probabilità

$$P_i = \frac{(170 - 110)}{455} = \frac{60}{455} = 0,132$$

# La tecnica di Lahiri

Vengono estratti due numeri casuali  $i$  e  $j$ :

$$1 \leq i \leq N \quad \text{e} \quad 1 \leq j \leq X_m$$

dove  $X_m = \max\{X_i\}$

Se  $j \leq X_i$  viene scelta l'unità  $i$

Se  $j > X_i$  l'unità corrispondente non viene scelta e viene estratta un'altra coppia di numeri casuali, ripetendo il confronto tra  $j$  e  $X_i$ .

Il processo di estrazione continua fino a quando non sia stata selezionata una unità.

La probabilità che si includa l'unità estratta già al primo tentativo è ( $X$  è l'ampiezza totale):

$$P_i = \sum_{i=1}^N \frac{1}{N} \frac{X_i}{X_m} = \frac{1}{NX_m} \sum_{i=1}^N X_i = \frac{X}{NX_m}$$

per ogni possibile unità contenuta nella popolazione, occorre moltiplicare la probabilità di essere scelta con il primo numero casuale per la probabilità che essa non venga rifiutata cioè

$$X_i / X_m$$

in quanto  $j$  porta ad accettare l'unità se si colloca nell'intervallo di numeri interi  $[1, X_i]$   
e a rifiutare se invece sta in  $[X_i + 1, X_m]$

Tale probabilità dipende da  $X_m$ . Se l'ampiezza massima è molto elevata, si devono prevedere più iterazioni per ottenere una unità accettabile

L'unica operazione da compiere prima di effettuare l'estrazione è la ricerca del valore massimo che le misure di ampiezza assumono sulla popolazione studiata!

# La tecnica di Lahiri

Vengono estratti due numeri casuali  $i$  e  $j$ :

dove  $1 \leq i \leq N$  e  $1 \leq j \leq X_m$   
 $X_m = \max\{X_i\}$

Se  $j \leq X_i$  viene scelta l'unità  $i$

Se  $j > X_i$  l'unità corrispondente non viene scelta e viene estratta un'altra coppia di numeri casuali, ripetendo il confronto tra  $j$  e  $X_i$ . Il processo di estrazione continua fino a quando non sia stata selezionata una unità.

La probabilità che si includa l'unità estratta già al primo tentativo è ( $X$  è l'ampiezza totale):

$$P_i = \sum_{i=1}^N \frac{1}{N} \frac{X_i}{X_m} = \frac{1}{NX_m} \sum_{i=1}^N X_i = \frac{X}{NX_m}$$

per ogni possibile unità contenuta nella popolazione, occorre moltiplicare **la probabilità di essere scelta con il primo numero casuale per la probabilità che essa non venga rifiutata** cioè  $X_i / X_m$  in quanto  $j$  porta ad accettare se si colloca nell'intervallo di numeri interi  $[1, X_i]$  e a rifiutare se invece sta in  $[X_i + 1, X_m]$ . Tale probabilità dipende da  $X_m$ . Se l'ampiezza massima è molto elevata, si devono prevedere più iterazioni per ottenere una unità accettabile. L'unica operazione da compiere prima di effettuare l'estrazione è la ricerca del valore massimo che le misure di ampiezza assumono sulla popolazione studiata!

## Esempio

Unità	1	2	3	4	5	6
Variabili						
$Y_i$	1.8	2.1	2.4	3.0	3.5	3.7
$X_i$	45	65	60	80	85	120

$X_m = 120$

Se la prima coppia è  $i=4, j=97$ , l'unità viene rifiutata in quanto

$X_4=80$  e  $j=97 > X_4$

Se ripetendo otteniamo  $i=1, j=54$  si rifiuta ancora e si prosegue in quanto  $X_1=45$  e  $j=54 > X_1$

Se  $i=2, j=41$  la seconda unità della popolazione viene accettata ed inclusa nel campione.

La probabilità di accettare alla prima estrazione è

$$P_i = \frac{455}{(6 \times 120)} = 0,632$$

# La tecnica di Lahiri

$$P_i = \sum_{i=1}^N \frac{1}{N} \frac{X_i}{X_m} = \frac{1}{NX_m} \sum_{j=1}^N X_j = \frac{X}{NX_m} \quad \text{è la probabilità di accettare l'unità } i\text{-esima alla prima estrazione}$$

Si dimostra che la probabilità di ottenere una certa unità  $i$  è uguale alla sua misura di ampiezza normalizzata:

Sia:  $p_i = \frac{X_i}{NX_m}$  la probabilità di accettare l'unità  $i$ -esima se estratta

$$q = \frac{1}{N} \sum_{j=1}^N \left(1 - \frac{X_j}{X_m}\right) = \frac{1}{NX_m} \sum_{j=1}^N (X_m - X_j) = \frac{NX_m - X}{NX_m} \quad \text{è la probabilità di non accettare l'unità in una estrazione}$$

$$P_i = p_i + qp_i + q^2 p_i + \dots = p_i (1 + q + q^2 + \dots) = \frac{p_i}{1 - q} = \frac{\frac{X_i}{NX_m}}{1 - \frac{NX_m - X}{NX_m}} = \frac{\frac{X_i}{NX_m}}{\frac{X}{NX_m}} = \frac{X_i}{X} \quad \text{è la probabilità di accettare l'unità } i$$

Si possono calcolare le probabilità di accettare un'unità alla seconda (e successive) iterazioni:

$$\text{Pr(accettare in due iterazioni)} = \text{Pr(accettare alla prima iterazione)} + \text{Pr(rifiutare alla prima)} * \text{Pr(accettare alla seconda)}$$

$$q = \frac{720 - 455}{720} = 0,368$$

$$\text{quindi } 0,632 + (0,368) \times 0,632 = 0,8646$$

# Confronto fra le due tecniche

## la *tecnica dei totali cumulati*

- è molto onerosa quando  $N$  è elevato
- se si vuole effettuare più di una estrazione senza ripetizione, è necessario ricalcolare le ampiezze cumulate ogni volta

## la *tecnica di Lahiri*

- non necessita di alcun calcolo aggiuntivo. La seconda unità viene estratta tra le  $N-1$  rimanenti con criterio identico alla prima
- vi è il vantaggio di una minore onerosità di calcolo rispetto all'altra tecnica in cambio di un più lungo e dispendioso procedimento di estrazioni di numeri casuali

# Metodi di estrazione senza ripetizione di un campione di ampiezza $n > 1$

- Tecniche per *l'estrazione di campioni* di ampiezza  $n$  con probabilità proporzionali alla dimensione e senza ripetizione
  - obiettivi dell'estrazione senza ripetizione
    - Con l'estrazione senza ripetizione si tende ad ottenere stime più precise rispetto a quelle che un campione della stessa ampiezza, ma estratto con ripetizione, può produrre
  - criteri guida con cui giudicare la convenienza di seguire l'uno anziché l'altro procedimento (PPS o PPC)
    - essi sono basati sulle proprietà delle probabilità di inclusione (primo e secondo ordine) che questi producono, dalle quali dipende la precisione e, in generale, la qualità degli stimatori



# Metodi di estrazione senza ripetizione di un campione di ampiezza $n > 1$

La scelta del metodo da utilizzare per estrarre campioni di dimensione  $n > 1$  si basa fondamentalmente sulle probabilità di inclusione di primo e di secondo ordine da cui dipende la precisione degli stimatori.

Le proprietà proposte da Hanurav che è bene abbiano le probabilità di inclusione sono:

i. *proporzionalità rispetto alle misure di ampiezza  $X_i$*

$$\pi_i = n(X_i / X) = \propto P_i \quad \forall i$$

indica il 'peso' di ogni unità: la probabilità di selezionare una unità è tanto maggiore quanto maggiore è il peso. È proprio questo il significato della prima condizione di Hanurav: l'unità  $i$  avrà una probabilità di essere inclusa nel campione proporzionale alla sua 'importanza', determinata dalla misura di ampiezza  $X_i$ .

sotto tale condizione, se si ha a disposizione una buona variabile ausiliaria  $X_i$  che approssimi la variabile oggetto di studio, lo stimatore del totale restituirà una stima con valori molto vicini al totale  $Y$  da stimare.

ii. *positività delle probabilità di inclusione del secondo ordine*

$$\pi_{ij} > 0 \quad \forall (i, j)$$

tale condizione è richiesta per l'esistenza di stimatori della varianza non distorti

# Metodi di estrazione senza ripetizione di un campione di ampiezza $n > 1$

Le proprietà delle probabilità di inclusione sono:

iii. *non negatività delle differenze*

tale condizione garantisce la non negatività della varianza dello stimatore

$$\pi_i \pi_j - \pi_{ij} \quad \forall (i, j)$$

iv. *soddisfacimento della relazione  $(\pi_{ij} / \pi_i \pi_j) > A$  per  $A > 0$  non prossimo allo zero*

tale condizione tende a stabilizzare le stime di varianza

# Metodi di estrazione senza ripetizione di un campione di ampiezza $n > 1$

I metodi più utilizzati sono:

- Il metodo di Yates e Grundy (1953)
- Il metodo di Brewer (1975)
- Il metodo di Sampford (1967)
- Il metodo di Rao, Hartley e Cochran (1962)
- Il metodo sistematico casualizzato (Madow 1949; Hartley 1966)

# Il metodo di Yates e Grundy (1953)

La prima unità viene estratta con probabilità proporzionale alle misure di ampiezza  $X_i$

$$P_i = X_i / X \quad \text{con} \quad X = \sum_{i=1}^N X_i$$

impiegando la tecnica dei totali cumulati o quella di Lahiri (estrazione ampiezza unitaria)

Per estrarre la seconda unità vengono modificate le probabilità di estrazione delle  $N-1$  unità residue. Se l'unità estratta è la  $i$ -esima, la probabilità di estrarre la seconda unità sarà:

$$P'_j = \frac{X_j}{X - X_i} = \frac{P_j}{1 - P_i} \quad j = 1, 2, \dots, i-1, i+1, \dots, N$$

La procedura può essere ripetuta in modo analogo per le successive unità da estrarre

■  $n=2$  la probabilità di inclusione del primo ordine dell'unità  $i$ -esima inclusa nel campione di due elementi è

$$\pi_i = P_i \left( 1 + \sum_{\substack{j=1 \\ j \neq i}}^N \frac{P_j}{1 - P_j} \right)$$

# Il metodo di Yates e Grundy (1953)

mentre la probabilità che le unità  $i$ -esima e  $j$ -esima siano incluse nel campione è:

$$\pi_{ij} = P_i \frac{P_j}{1 - P_i} + P_j \frac{P_i}{1 - P_j}$$

*probabilità di inclusione del secondo ordine*

■  $n=3$  le probabilità di inclusione, rispettivamente del primo e del secondo ordine sono:

$$\pi_i = P_i \left( 1 + \sum_{\substack{j=1 \\ j \neq i}}^N \frac{P_j}{1 - P_j} + \sum_{\substack{j=1 \\ j \neq i}}^N \frac{P_j}{1 - P_j} \sum_{\substack{K=1 \\ K \neq i, j}}^N \frac{P_K}{1 - P_j - P_K} \right)$$

$$\pi_{ij} = P_i P_j \left[ \frac{1}{1 - P_i} \left( 1 + \sum_{\substack{k=1 \\ k \neq i, j}}^N \frac{P_k}{1 - P_j - P_k} \right) + \frac{1}{1 - P_j} \times \left( 1 + \sum_{\substack{k=1 \\ k \neq i, j}}^N \frac{P_k}{1 - P_j - P_k} \right) + \sum_{\substack{k=1 \\ k \neq i, j}}^N \frac{P_k}{1 - P_k} \left( \frac{P_k}{1 - P_i - P_k} + \frac{P_k}{1 - P_j - P_k} \right) \right]$$

$\frac{1}{1 - P_k}$  non  $\frac{P_k}{1 - P_k}$   
 $\frac{P_k}{1 - P_j - P_k}$  non  $\frac{P_k}{1 - P_j}$

con queste correzioni la formula ti dà il risultato corretto

# Il metodo di Yates e Grundy (1953)

## Esempio

Estraiamo un campione di numerosità  $n=3$ .

Le probabilità di estrazione devono essere proporzionali alle ampiezze, cioè a valori noti di  $X$ .

Unità	1	2	3	4	5	6
Variabili						
$Y_i$	1.8	2.1	2.4	3.0	3.5	3.7
$X_i$	45	65	60	80	85	120
$P_i=X_i/X$	0,099	0,143	0,132	0,176	0,187	0,264

La probabilità di inclusione nel campione della prima unità della popolazione è:

$$\pi_1 = P_1 \left( 1 + \sum_{j=2}^6 \frac{P_j}{1 - P_j} + \sum_{j=2}^6 \frac{P_j}{1 - P_j} \sum_{\substack{k=2 \\ k \neq j}}^6 \frac{P_k}{1 - P_j - P_k} \right) =$$

$$0,099 \times \left[ 1 + \frac{0,143}{1 - 0,143} + \frac{0,132}{1 - 0,132} + \frac{0,176}{1 - 0,176} + \frac{0,187}{1 - 0,187} + \frac{0,264}{1 - 0,264} + \frac{0,143}{1 - 0,143} \left( \frac{0,132}{1 - 0,143 - 0,132} + \frac{0,176}{1 - 0,143 - 0,176} + \frac{0,187}{1 - 0,143 - 0,187} + \frac{0,264}{1 - 0,143 - 0,264} \right) + \dots \right] = 0,337$$

con calcoli analoghi si ottengono le altre probabilità di inclusione del primo ordine

Unità	1	2	3	4	5	6
$\pi_i$	0.337	0.457	0.429	0.534	0.557	0.686

# Il metodo di Yates e Grundy (1953)

Tramite calcoli ancora più laboriosi si ottengono le probabilità di inclusione del secondo ordine.

È facile verificare che la somma delle probabilità di inclusione del primo ordine per estrazione senza ripetizione è pari a  $n$  (altra proprietà di base per estrazioni senza ripetizione...)

$$\sum_{i=1}^N \pi_i = n$$

Infatti la somma delle probabilità riportata nella tabella seguente è pari a 3.

Unità	1	2	3	4	5	6
Prob.						
$3P_i$	0.297	0.429	0.369	0.528	0.561	0.792
$\pi_i$	0.337	0.457	0.429	0.534	0.557	0.686

per questo metodo invece non vale la proprietà di proporzionalità tra le probabilità di inclusione del primo ordine  $\pi_i$  e le probabilità iniziali di estrazione  $\pi_i = nP_i$  che nel caso in esame dovrebbero essere  $\pi_i = 3P_i$ .

# Il metodo di Brewer (1975)

Esamineremo il caso di campioni di dimensione  $n=2$

Il procedimento può essere così schematizzato:

- Si estrae una prima unità con probabilità  $\frac{1}{D} \frac{P_i(1-P_i)}{1-2P_i}$ ,  $i = 1, 2, \dots, N$

dove  $D$  è un fattore di normalizzazione  $D = \sum_{i=1}^N \frac{P_i(1-P_i)}{1-2P_i}$   $j \neq i$

Ammesso che nella prima estrazione sia stata selezionata l'unità  $i$  si estrae una seconda unità con probabilità

$$P_j / (1 - P_i)$$

*In pratica la prima unità viene estratta con probabilità modificata rispetto alle semplici misure di ampiezza normalizzate  $P_i$ ; la seconda estrazione, invece, impiega probabilità uguali a quelle del metodo di Yates and Grundy.*



# Il metodo di Brewer (1975)

*la probabilità di inclusione del primo ordine si calcola in questo modo:*

$$\pi_i = \frac{P_i(1-P_i)}{D(1-2P_i)} + \sum_{j \neq i}^N \frac{P_j(1-P_j)}{D(1-2P_j)} \frac{P_i}{(1-P_j)} = 2P_i$$

*la probabilità di inclusione del secondo ordine è pari a:*

$$\pi_{ij} = \frac{P_i P_j}{D} \left( \frac{1}{1-2P_i} + \frac{1}{1-2P_j} \right) \quad j \neq i$$

Le probabilità con cui sono state estratte le due unità sono tali che soddisfano le prime tre proprietà formulate da Hanurav:

1. la probabilità di inclusione di primo ordine soddisfa la relazione  $\pi_i = 2P_i$
2.  $\pi_{ij} > 0$
3.  $\pi_{ij} - \pi_i \pi_j > 0$  per ogni  $i$  e  $j$

# Il metodo di Brewer (1975)

## Esempio

N=8 Comuni di un comprensorio montano

Y= fatturato annuo del settore turistico

X= variabile ausiliare data dal n° di esercizi alberghieri per comune

Unità	1	2	3	4	5	6	7	8	
Y <sub>i</sub>	0.90	1.29	0.02	1.70	2.27	1.12	1.90	2.49	
X <sub>i</sub>	6	5	1	7	10	7	9	12	57

si vuole estrarre un campione di n=2 Comuni mediante il metodo di Brewer  
la probabilità modificata per la prima estrazione è (con D=1,224):

$$\frac{0,105(1-0,105)}{(1-2 \times 0,105)} \frac{1}{1,224} = 0,097$$

In tabella le probabilità di estrazione delle altre unità:

Unità	1	2	3	4	5	6	7	8
Pr. mod.	0.097	0.079	0.015	0.117	0.182	0.117	0.159	0.235
P <sub>i</sub>	0.105	0.088	0.018	0.123	0.175	0.123	0.158	0.211

# Il metodo di Brewer (1975)

## Esempio

rispetto alle  $P_i$  i valori più elevati risultano ulteriormente incrementati e quelli meno elevati risultano ulteriormente ridotti

Unità	1	2	3	4	5	6	7	8
Pr. mod.	0.097	0.079	0.015	0.117	0.182	0.117	0.159	0.235
$P_i$	0.105	0.088	0.018	0.123	0.175	0.123	0.158	0.211

La probabilità di inclusione del primo ordine è:

$$\begin{aligned}\pi_i &= \frac{P_i(1-P_i)}{D(1-2P_i)} + \sum_{j \neq i} \frac{P_j(1-P_j)}{D(1-2P_j)} \frac{P_i}{(1-P_j)} = 2P_i \\ \pi_i &= \frac{P_i(1-P_i)}{D(1-2P_i)} + \sum_{j \neq i} \frac{P_j(1-P_j)}{D(1-2P_j)} \frac{P_i}{(1-P_j)} = \\ &= \frac{0,105(1-0,105)}{1,224(1-2 \times 0,105)} + [(0,115 \times 0,079) + \dots] = 0,097 + 0,113 = 0,21 \\ \pi_i &= 2P_i = 2 \times 0,105 = 0,21\end{aligned}$$

# Il metodo di Brewer (1975)

La probabilità di ottenere una qualsiasi seconda unità data la prima sono riassunte di seguito:

Seconda	1	2	3	4	5	6	7	8
Prima								
1	-	0.098	0.020	0.137				
2	0.115	-						
3	0.107	0.089	-					
4				-				
5					-			
6						-		
7							-	
8								-

la probabilità di ottenere come seconda unità campionaria la prima unità della popolazione, dopo aver già estratto la seconda unità della popolazione è:

$$(P_1/(1-P_2)=0,105/(1-0,088)=0,115$$

*da notare che tali valori non sono simmetrici ( $P_2/(1-P_1)=0,098$*

# Il metodo di Sampford (1967)

Tecnica diversa da quelle precedentemente illustrate perché consiste:

- A. nell'estrazione con ripetizione e con probabilità assegnate di  $n$  unità**
- B. nel rifiuto del campione ove qualche unità si presenti più di una volta**

Procedimento:

- si estrae una prima unità con probabilità  $P_i = X_i/X$
- le successive unità vengono estratte con probabilità proporzionali a  $P_i/(1-nP_i)$  e con fattore di proporzionalità  $1/D$
- l'estrazione avviene con ripetizione

*N.B. Se una unità è estratta due volte, si rifiuta il campione estratto e si ricomincia il procedimento*

# Il metodo di Sampford (1967)

Per calcolare la probabilità di inclusione di primo ordine si può procedere nel modo seguente:

$$\pi_i = \sum_{k=0}^{\infty} \Pr(\text{estrarre } k \text{ campioni con qualche elemento ripetuto, che vengono rigettati}) \times$$

$\times \Pr(\text{estrarre al } k + 1 \text{esimo tentativo un campione di elementi distinti di cui uno è l}'i\text{-esimo})$

in particolare se  $n=2$  è pari a:

$$\pi_i = P_i \left( 1 + \frac{1}{D} \right) \sum_{k=0}^{\infty} \left( \sum_{j=1}^N P_j \frac{1 - 2P_j}{D} \right)^k$$

la probabilità di inclusione del secondo ordine è:

$$\pi_{ij} = \frac{P_i(1-P_i)}{D(1-2P_i)} \frac{P_j}{1-P_i} + \frac{1}{D} \frac{P_j(1-P_j)}{1-2P_j} \frac{P_i}{1-P_j} = \frac{P_i P_j}{D} \left( \frac{1}{1-2P_i} + \frac{1}{1-2P_j} \right)$$

***Il calcolo delle probabilità di ordine superiore è molto complesso***

# Il metodo di Sampford (1967)

## Esempio

$n=3$

per la prima estrazione le probabilità sono le  $P_i$  già calcolate con il metodo di YG.

La probabilità per le successive unità (dopo la prima estrazione) è pari a:

$$\frac{P_i}{1 - nP_i} \frac{1}{D} = \frac{0,099}{1 - 3 \cdot 0,099} \frac{1}{2,669} = 0,053$$

La tabella riporta le probabilità di tutte le unità della popolazione delle estrazioni successive alla prima in base al metodo di Sampford.

Unità	1	2	3	4	5	6
Variabili						
$P_i$	0.099	0.143	0.132	0.176	0.187	0.264
Pr. estrazione successiva	0.053	0.094	0.082	0.139	0.159	0.473

# Il metodo di Sampford (1967)

## Esempio

Una applicazione del metodo di Sampford può essere la seguente:

### ■ Passo 1:

□ Si estrae con tale procedura il campione ordinato di 3 unità: (2, 6, 2), con probabilità  $0,143 \times 0,473 \times 0,094 = 0,0063$ . Dato che l'unità 2 è ripetuta si rifiuta il campione e si procede a una nuova estrazione;

### ■ Passo 2:

□ Si estrae il campione (4, 4, 2) con probabilità  $0,176 \times 0,139 \times 0,094 = 0,0023$ . Rifiutiamo il campione per lo stesso motivo del passo 1:

### ■ Passo 3:

□ Si estrae un terzo campione ordinato (1, 5, 6) con probabilità  $0,099 \times 0,159 \times 0,473 = 0,0074$  che viene accettato perché nessuna unità è ripetuta

Con tale metodo la probabilità di inclusione del primo ordine è proporzionale alle misure d'ampiezza  $\pi_i = nP_i$  e la probabilità di inclusione del secondo ordine è positiva.

Unità	1	2	3	4	5	6
$\pi_i = 3P_i$	0.297	0.429	0.369	0.528	0.561	0.792



# Il metodo di Rao, Hartley e Cochran (1962)

Con questa tecnica le unità campionarie non vengono selezionate direttamente dalla popolazione di riferimento, ma da sottopopolazioni costruite prima.

- La popolazione viene suddivisa casualmente in  $n$  gruppi di numerosità  $N_1, N_2, \dots, N_n$ , tali che

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

- Calcolata per ogni gruppo la misura di ampiezza totale

$$X_{h.} = \sum_{j=1}^{N_h} X_{hj}, h = 1, 2, \dots, n$$

si estrae dal gruppo stesso una unità con probabilità

$$P_{hj} = X_{hj} / X_{h.}, j = 1, 2, \dots, N_h$$

Il procedimento per calcolare la probabilità di inclusione  $\pi_i$  per la generica unità  $i$ -esima risulta piuttosto complesso .....

# Il metodo di Rao, Hartley e Cochran (1962)

Calcolo della probabilità di inclusione del primo ordine  $\pi_i$  per la generica unità  $i$

- A. Occorre sommare la probabilità che tale unità sia inclusa in ciascun gruppo
- B. moltiplicarla per la probabilità di estrarla dal gruppo medesimo.

Poiché quest'ultima dipende non solo dalla sua misura di ampiezza, ma anche dall'ampiezza totale del gruppo  $X_h$



è necessario considerare tutte le possibili ampiezze dei gruppi che possono contenere l'unità  $i$ -esima.

*Si dimostra che le probabilità di inclusione del primo ordine non sono proporzionali a  $P_i$*

# Il metodo di Rao, Hartley e Cochran (1962)

## Esempio:

Consideriamo sempre le  $N=6$  famiglie ed estraiamo un campione di  $n=3$  con il metodo RHC. Suddividiamo la popolazione in  $n=3$  gruppi, con numerosità  $N_1=N_2=N_3=2$

Gruppi	G1	G2	G3
Unità nei gruppi	1;4	2;3	5;6
$X$	45,80	65,60	85,120
$X_h$	125	125	205

assegnata casualmente l'appartenenza di ogni unità a una delle tre coppie, si procede ad una estrazione da ogni coppia.

Supponiamo di estrarre casualmente le unità (4,2,6)

la probabilità sono  $P_{1,2}=80/125= 0,64$

$$P_{2,1}=65/125= 0,52$$

$$P_{3,2}=120/205= 0,59 \text{ (il primo pedice individua il gruppo, il secondo l'unità)}$$

# Il metodo sistematico casualizzato (Madow 1949; Hartley 1966)

La tecnica può essere schematizzata come segue:

- Le N unità della popolazione vengono elencate e numerate in modo casuale

- Vengono determinate le quantità  $X'_i = nX_i$  dove n è un'opportuna dimensione campionaria fissata

- Vengono individuate le corrispondenti quantità cumulate

$$T'_i = \sum_{j=1}^i X'_j$$

- Si estrae un numero casuale r  $1 \leq r \leq X$

- Le unità campionarie estratte saranno quelle cui corrispondono valori cumulati immediatamente superiori alle quantità r, r+X, r+2X, ...r+(n-1)X

***le probabilità di inclusione del primo ordine  
sono proporzionali alle ampiezze normalizzate***

# Il metodo sistematico casualizzato (Madow 1949; Hartley 1966)

È un metodo facilmente adattabile per campioni medio-grandi.  
Le unità devono essere ordinate preventivamente in modo casuale

## Esempio:

Si vuole estrarre un campione di dimensione  $n=3$  supponendo che l'estrazione del numero casuale sia  $r=15$  [1,57]. Le unità della popolazione che formano il campione sono quelle corrispondenti ai valori  $T_i$  immediatamente superiori a

$57 = X$ , cioè il totale cumulato della variabile  $X_i$

$r=15;$ $r+X=15+57=72$ $r+2X=15+114=129$	2	Unità	3	2	7	4	8	5	1	6
		$Y_i$	0.02	1.29	1.90	1.70	2.49	2.27	0.90	1.12
		$X_i$	1	5	9	7	12	10	6	7
		$3X_i$	3	15	27	21	36	30	18	21
		$T_i$	3	18	45	66	102	132	150	171

praticamente la prima unità estratta è la prima superiore a 15 (con  $T_i$ ), dove  $T_i$  è la cumulata di  $3X_i$ ,  $3=n$  ampiezza campionaria  
per la seconda unità si considera quella unità che ha  $T_i$  appena maggiore di  $r + 2X_i$  e così via

Le unità estratte saranno: 2 – 8 – 5

# Metodi di estrazione con ripetizione di un campione di ampiezza $n > 1$

Per l'estrazione con probabilità proporzionale all'ampiezza e con ripetizione di  $n$  unità si tratta di ripetere  $n$  volte il procedimento descritto per il caso di un campione unitario.

Le probabilità di inclusione del primo e del secondo ordine sono date da:

$$\pi_i = 1 - (1 - P_i)^n$$

$$\pi_{ij} = 1 - (1 - P_i)^n - (1 - P_j)^n + (1 - P_i - P_j)^n$$

Esempio:

Data una popolazione di 10 classi liceali si vuole estrarre un campione di tre classi con probabilità proporzionale alla dimensione e con ripetizione.

Dalla tavola dei numeri casuali vengono estratti i numeri 30, 107 e 164, quindi, adoperando la tecnica dei totali cumulati le classi scelte saranno 2,6,10

Classe	Ampiezza	Totali cumulati
1	15	15
2	20	35
3	18	53
4	12	65
5	25	90
6	21	111
7	20	131
8	14	145
9	15	160
10	17	177

# Stima del totale

## Campioni estratti con ripetizione

Si consideri un piano di campionamento probabilistico che preveda l'estrazione con ripetizione di campioni di ampiezza fissa  $n$ .

Siano

- $y_1, y_2, \dots, y_n$  le osservazioni campionarie
- $p_1, p_2, \dots, p_n$  le probabilità di estrazione delle unità a cui, nell'ordine, le osservazioni si riferiscono
- uno stimatore non distorto del totale  $\hat{Y}$  è dato da:

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i}$$

Esso è noto come stimatore di Hansen-Hurwitz.

È una combinazione lineare delle osservazioni campionarie con pesi  $1/np_i$  ( $i=1, 2, \dots, n$ ) che dipendono dalle etichette delle unità a cui si riferiscono.

# Stima del totale

## Campioni estratti con ripetizione

La varianza dello stimatore assume la forma

$$V(\hat{Y}) = \frac{1}{n} \sum_{i=1}^N P_i \left( \frac{Y_i}{P_i} - Y \right)^2 \quad (*)$$

e può essere stimata correttamente con

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

$$V(\hat{Y}_{HH}) = 0 \quad \text{quando} \quad P_i = \frac{X_i}{X} = \frac{Y_i}{Y} \quad \forall i$$

***cioè quando sussiste una perfetta proporzionalità tra la variabile ausiliare X e la variabile oggetto di studio Y***

Nella pratica è sufficiente una buona proporzionalità fra  $X_i$  e  $Y_i$  per avere il vantaggio di usare probabilità variabili e varianze degli stimatori più contenute.



# Stima del totale

## Campioni estratti senza ripetizione

Si consideri un piano di campionamento probabilistico che preveda l'estrazione senza ripetizione di campioni di ampiezza non necessariamente fissa.

Siano  $y_1, y_2, \dots, y_n$  le osservazioni campionarie e  $\pi_1, \pi_2, \dots, \pi_n$  le probabilità di inclusione delle unità della popolazione a cui, nell'ordine, le osservazioni si riferiscono. Allora uno stimatore non distorto del totale è dato da:

$$\hat{Y} = \sum_{i=1}^n \frac{y_i}{\pi_i}$$

Esso è noto come lo stimatore HT di Horvitz-Thompson

La varianza dello stimatore assume la forma

$$V(\hat{Y}) = \sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} Y_i^2 + \sum_{i=1}^N \sum_{j \neq i}^N \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) Y_i Y_j$$

# Stima del totale

## Campioni estratti senza ripetizione

se le probabilità di inclusione del secondo ordine sono tutte positive, può essere stimata correttamente con

$$\hat{v}(\hat{Y}_{HT}) = \sum_{i=1}^n \frac{1-\pi_i}{\pi_i^2} y_i^2 + \sum_{i=1}^n \sum_{j \neq i}^n \left( \frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}} \right) y_i y_j$$

Lo stimatore HT è funzione delle probabilità del primo ordine mentre la varianza dello stimatore è funzione della probabilità di inclusione di primo e secondo ordine.

# Stima del totale

## Campioni estratti senza ripetizione

Si consideri un piano di campionamento probabilistico che preveda l'estrazione senza ripetizione di campioni di ampiezza  $n$  fissata.

Lo stimatore del totale di HT vale

La varianza dello stimatore assume la seguente forma:

$$V'(\hat{Y}_{HT}) = \sum_{i=1}^N \sum_{j>i}^N (\pi_i \pi_j - \pi_{ij}) \left( \frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2$$

se le probabilità di inclusione del secondo ordine sono tutte positive, può essere stimata correttamente con lo stimatore di Yates e Grundy

$$\hat{v}'(\hat{Y}_{HT}) = \sum_{i=1}^n \sum_{j>i}^n \left( \frac{\pi_i \pi_j}{\pi_{ij}} - 1 \right) \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

# Stima della media

- Sia  $T$  un qualsiasi stimatore del totale della popolazione  $Y$
- Sia  $E(T)$  la media di  $T$
- Sia  $V(T)$  e la varianza di  $T$ .

la quantità  $\bar{T} = T / N$  è uno stimatore di  $\bar{Y}$  con media e varianza pari, rispettivamente, a

$$E(\bar{T}) = \frac{E(T)}{N}, \quad V(\bar{T}) = \frac{V(T)}{N^2}$$