Tecniche di indagine statistica

Lezione 16/17

Campionamento probabilistico – concetti base

Popolazione *finita* di N unità U_i con $i = 1, 2, \ldots, N$

Variabile: Y $(Y_1, Y_2, \ldots, Y_N)$ **n.b.: $Y_1, Y_2, \ldots, Y_N$ costanti**

Caratteristiche di Y *(parametri della popolazione finita)*:

totale: $t = \sum_{i=1}^{N} Y_i$

media: $\overline{Y}_U = \frac{1}{N} \sum_{i=1}^{N} Y_i = \frac{t}{N}$

proporzione:

$P = \sum_{i=1}^{N} \frac{Y_i}{N}$ con

$Y_i = \begin{cases} 1 & \text{se l'unità } i\text{-esima ha la caratteristica d'interesse} \\ 0 & \text{altrimenti} \end{cases}$

Var (Y): $S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \overline{Y}_U)^2$

Inferenza sui parametri di Y: selezione campione di n elementi da U

Selezione campione di *n* elementi da *U*

Caratteristiche del campione:

- elementi (unità) di *U*: oggetti 'reali'
  preferibile avere unità diverse

- non interessa il 'momento' (ordine) in cui l'elemento è inserito nel campione

- da una pop.ne *U* di dimensione N, gli insiemi di *n* con queste caratteristiche (campioni di dimensione *n)* sono: $\binom{N}{n}$

Campionamento probabilistico:

1. ogni possibile campione ha una probabilità nota $p(c)$ di essere scelto
2. da cui può essere calcolata la probabilità che una specifica unità *k* sia inclusa nel campione

Campionamento casuale semplice: $p(c)$ uguale $\forall\ c$

Campionamento Casuale Semplice (CCS)*

Disegno di campionamento *base*. Poiché ci sono $\binom{N}{n}$ campioni: uguale probabilità di selezione ad ogni campione **c** di ampiezza ***n***

$$p(c) = \frac{1}{\binom{N}{n}} \qquad \forall c$$

I_k *variabile indicatrice* della presenza dell'unità *k* in **c**:

I_k = *1 se unità k in c*

= *0 altrimenti*

I_k variabile casuale -bernoulliana- che dipende da $p(c)$

probabilità di inclusione unità *k* $\pi_k = \Pr(I_k = 1) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$

$$E(I_k) = \pi_k$$
$$V(I_k) = \pi_k (1 - \pi_k)$$

campioni che includono unità *k*
di possibili campioni

- La probabilità di inclusione è uguale alla frazione di campionamento $\frac{n}{N}$ per ogni unità

**** senza ripetizione***

Campionamento Casuale Semplice (CCS)

Campionamento *senza reinserimento*: unità *k* inclusa nel campione solo una volta

Stima parametri popolazione finita:

media $\bar{y} = \frac{1}{n}\sum_{i \in c} Y_i = \frac{1}{n}\sum_{k=1}^{N} Y_k I_k$

I_k = *1 se unità k in c*
= *0 altrimenti*

Stimatore ***non distorto*** $E(\bar{y}) = E\left(\sum_{k=1}^{N} \frac{Y_k I_k}{n}\right) = \frac{1}{n}\sum_{k=1}^{N} Y_k E(I_k)$

$$E(\bar{y}) = \frac{1}{n}\sum_{k=1}^{N} Y_k \frac{n}{N} = \overline{Y}_U$$

Varianza stimatore media campionaria - CCS

$$V(\overline{y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}$$

S^2 in genere ***non nota***

$$\hat{V}(\overline{y}) = (1-f)\frac{s^2}{n}$$

$$s^2 = \frac{1}{n-1}\sum_{i \in c}(Y_i - \overline{y})^2$$

$$\widehat{SE}(\overline{y}) = \sqrt{\hat{V}(\overline{y})}$$

$$\left(1 - \tfrac{n}{N}\right) = (1-f)$$

fattore di correzione per pop.ni finite

(f = n/N frazione di campionamento)

Varianza stimatore media campionaria – CCS (da pop.ne finita)

$$
\begin{aligned}
V(\bar{y}) &= \frac{1}{n^2} V\left(\sum_{i=1}^{N} Z_i y_i\right) \\
&= \frac{1}{n^2} \mathrm{Cov}\left(\sum_{i=1}^{N} Z_i y_i, \sum_{j=1}^{N} Z_j y_j\right) \\
&= \frac{1}{n^2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \,\mathrm{Cov}\,(Z_i, Z_j) \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 V(Z_i) + \sum_{i=1}^{N} \sum_{j \neq i}^{N} y_i y_j \,\mathrm{Cov}\,(Z_i, Z_j)\right] \\
&= \frac{1}{n^2}\left[\frac{n}{N}\left(1-\frac{n}{N}\right)\sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N}\sum_{j\neq i}^{N} y_i y_j \frac{1}{N-1}\left(1-\frac{n}{N}\right)\left(\frac{n}{N}\right)\right] \\
&= \frac{1}{n^2}\frac{n}{N}\left(1-\frac{n}{N}\right)\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j\neq i}^{N} y_i y_j\right] \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N(N-1)}\left[(N-1)\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2 + \sum_{i=1}^{N} y_i^2\right] \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N(N-1)}\left[N\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2\right] \\
&= \left(1-\frac{n}{N}\right)\frac{S^2}{n}.
\end{aligned}
$$

Z_i variabile indicatrice presenza unità *i* nel campione *c*

Intervalli di confidenza per media campionaria - CCS

$$\Pr(l(\hat{\theta}) \leq \theta \leq u(\hat{\theta})) = 1 - \alpha$$

Ipotizzando che:

- ci sia una sequenza di pop.ni finite di dimensione *N* crescente
- per ciascuna pop.ne si ha un campione di dimensione *n* che cresce al crescere di *N*
- *N* – *n* diventi arbitrariamente grande

Formulazione del teorema del limite centrale che consente di approssimare: $\frac{\overline{y} - \overline{Y}_u}{SE(\overline{y})} = \frac{\overline{y} - \overline{Y}_u}{\sqrt{1-f}\frac{s}{\sqrt{n}}} \rightarrow N(0,1)$

Per cui IC a livello *(1-α)* per media campionaria in CCS:

$$\overline{y} - z_{\alpha/2} SE(\overline{y});\ \overline{y} + z_{\alpha/2} SE(\overline{y})$$

con $z_{\alpha/2}$ percentile di livello *(1-α/2)* da *N* (0,1)

Campionamento Casuale Semplice (CCS)

Stima parametri popolazione finita:

totale $\hat{t} = N\,\overline{y}$ con $V(\hat{t}) = N^2\,V(\overline{y})$

stimata da $\hat{V}(\hat{t}) = N^2(1-f)\dfrac{s^2}{n}$

Proporzione, poiché $P = \sum_{i=1}^{N} \dfrac{Y_i}{N}$ con

$$Y_i = \begin{cases} 1 & \text{se l'unità } i\text{-esima ha la caratteristica d'interesse} \\ 0 & \text{altrimenti} \end{cases}$$

$$\hat{p} = \sum_{k=1}^{N} \frac{Y_k\,I_k}{n} = \overline{y}$$ Y dicotomica $S^2 = \dfrac{N}{N-1}\,p\,(1-p)$

$$V(\hat{p}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right) = \left(\frac{N-n}{N-1}\right)\frac{p\,(1-p)}{n}$$

data la non distorsione di $\overline{y}$, anche non distorsione di $\hat{t}$ e $\hat{p}$.

Stima varianza *proporzione* campionaria in CCS

$$V(\hat{p}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right) = \left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n}$$

stimata da

$$s^2 = \frac{1}{n-1}\sum (y_i - \hat{p})^2 = \frac{n}{n-1}\hat{p}(1-\hat{p}).$$

$$\hat{V}[\hat{p}] = \left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}$$

Campione casuale semplice (CCS) – rappr.ne grafica

senza reinserimento (in blocco, senza ripetizione)

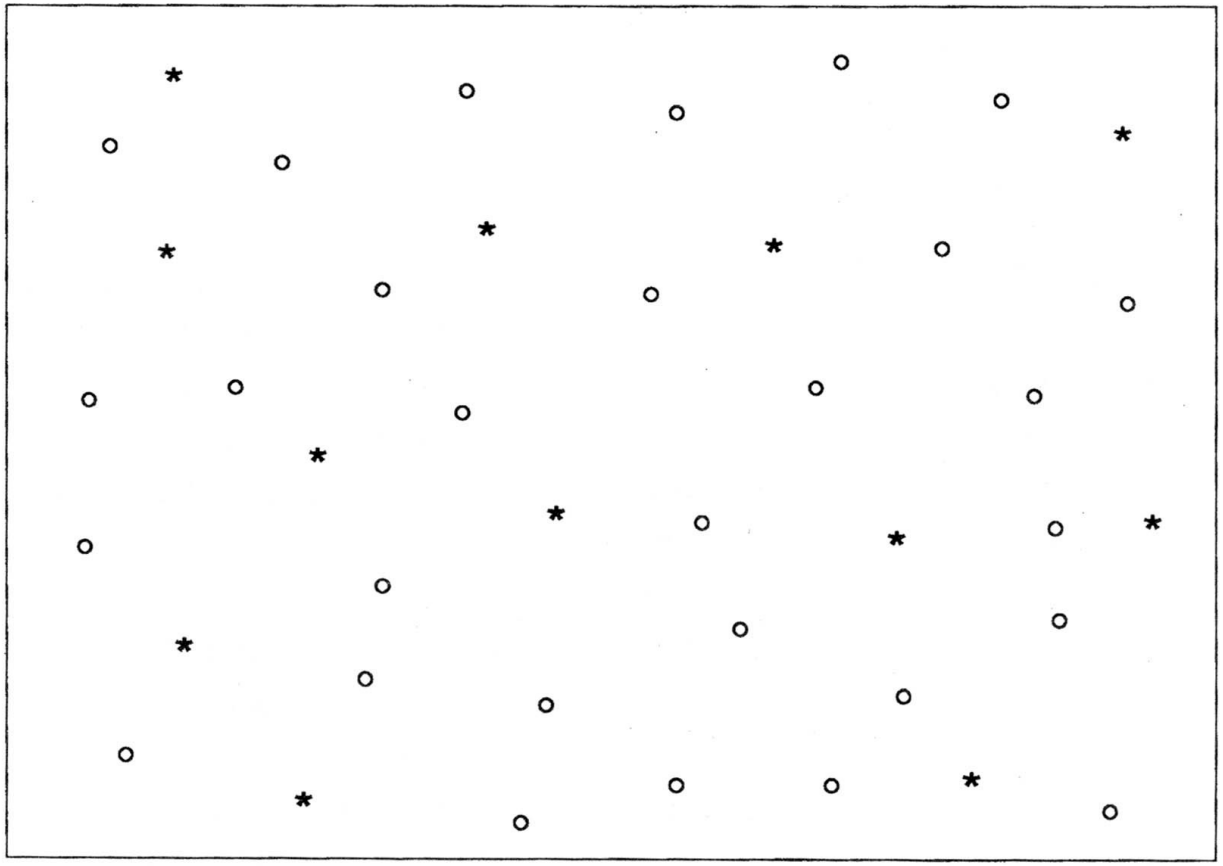

Fig. 1. Campionamento casuale semplice: unità della popolazione (∘), unità campionaria (⋆).

Esempio CCS

Popolazione di N = 8 unità ($U_1, U_2, \ldots, U_8$)

CCS di n = 4 unità per stimare t (= 40)

70 possibili campioni di ampiezza 4, ciascuno con $p(c)$ = 1/70

$$\{Y_i\} = \{1, 2, 4, 4, 7, 7, 7, 8\}$$

ovvero

$$Y_2 = 2, \quad Y_3 = 4, \ldots$$

estrarre un campione c significa considerare un sottoinsieme di etichette in U.

Sample S	$y_i, i \in S$	$\hat{t}_S$
{1, 2, 3, 4}	1, 2, 4, 4	22
{1, 2, 3, 5}	1, 2, 4, 7	28
{1, 2, 3, 6}	1, 2, 4, 7	28
{1, 2, 3, 7}	1, 2, 4, 7	28
{1, 2, 3, 8}	1, 2, 4, 8	30
{1, 2, 4, 5}	1, 2, 4, 7	28
{1, 2, 4, 6}	1, 2, 4, 7	28
{1, 2, 4, 7}	1, 2, 4, 7	28
{1, 2, 4, 8}	1, 2, 4, 8	30
{1, 2, 5, 6}	1, 2, 7, 7	34
⋮	⋮	⋮

Campione estratto:

$$c = \{2, 3, 6, 7\}$$

cui corrispondono i valori

$Y_2 = 2$, $Y_3 = 4$, $Y_6 = 7$ e $Y_7 = 7$, per cui

$$\hat{t}_c = N \frac{\sum_{k=1}^{N} Y_k I_k}{n}, \quad I_k = \begin{cases} 1 & \text{se } i \in c \\ 0 & \text{altrimenti} \end{cases}$$

= 40

Distribuzione campionaria stimatore di *t* - Esempio

k	22	28	30	32	34	36	38	40	42	44	46	48	50	52	58
$P(\hat{t} = k)$	$\frac{1}{70}$	$\frac{6}{70}$	$\frac{2}{70}$	$\frac{3}{70}$	$\frac{7}{70}$	$\frac{4}{70}$	$\frac{6}{70}$	$\frac{12}{70}$	$\frac{6}{70}$	$\frac{4}{70}$	$\frac{7}{70}$	$\frac{3}{70}$	$\frac{2}{70}$	$\frac{6}{70}$	$\frac{1}{70}$