Tecniche di indagine statistica

Lezione 18

Come selezionare un CCS

Procedura di selezione: garantire ad ogni unità in *N* la *stessa probabilità* di far parte del campione

1. Tavola di numeri casuali (generatori di n.ri peuso-casuali - funzione CASUALE() in Excel o procedura sample in R)

2. Selezione "sistematica"

Selezione CCS mediante numeri *casuali*

	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40
1	77 66	88 40	86 61	96 70	78 75	29 77	21 94	12 37	66 11	53 42
2	74 81	53 71	16 61	59 13	33 02	25 95	92 37	03 18	46 26	37 86
3	05 88	20 12	10 45	80 22	38 70	94 11	22 02	08 37	74 87	49 04
4	05 79	76 95	69 00	48 70	60 14	53 11	06 57	06 26	60 31	06 74
5	79 98	70 98	97 94	55 99	44 04	75 89	69 50	64 03	96 68	17 89
6	55 09	79 15	11 56	65 88	08 16	96 95	33 17	60 45	81 31	50 46
7	79 19	16 49	99 08	80 01	56 35	41 42	72 58	20 39	33 53	85 26
8	28 70	12 06	71 02	34 50	30 16	83 58	39 98	84 01	27 85	17 35
9	54 44	53 59	34 44	49 93	61 75	19 87	34 93	85 16	18 79	65 94
10	93 69	31 43	93 93	77 39	72 40	66 32	90 86	65 88	41 19	36 86
11	24 94	65 41	64 64	95 13	46 97	43 12	86 02	79 50	67 90	14 19
12	04 07	67 01	59 03	27 37	83 20	17 82	11 80	46 08	32 68	60 26
13	67 24	63 38	76 53	29 14	02 47	70 31	20 88	24 31	14 65	23 35
14	69 06	90 51	48 94	89 77	41 66	54 60	66 95	46 73	76 59	20 05
15	66 56	20 91	61 48	91 73	98 80	96 94	45 09	93 21	90 40	03 01
16	36 48	02 01	88 94	20 08	07 64	08 84	26 41	25 54	43 65	82 24
17	62 93	85 57	12 06	07 88	22 37	03 84	80 69	93 29	22 34	67 88
18	94 01	05 57	71 98	47 26	58 99	72 11	69 93	22 46	72 52	75 62
19	52 94	18 97	82 49	76 84	86 83	05 27	53 27	16 40	94 34	81 86
20	27 43	78 39	71 17	16 72	43 37	60 73	83 41	31 32	61 05	37 89
21	46 00	19 71	63 06	75 27	01 57	59 61	86 70	33 35	54 77	81 38
22	29 58	01 44	39 62	83 16	97 46	31 27	27 43	67 66	35 08	86 34
23	19 31	80 79	63 47	80 56	00 71	06 17	49 70	26 75	55 43	46 84
24	02 52	31 23	74 12	16 62	21 19	76 63	33 43	17 16	96 00	42 50
25	06 00	13 63	57 37	51 83	45 58	21 01	02 89	88 07	74 32	21 87

Fonte: M.G. Kendall e B. Babington Smith, *Tables of Random Sampling Numbers*, Cambridge University Press, 1954.

N = 750 (unità numerate da 001-750)

- Gruppi di 3 cifre (esclusi 000 e > 750):

776, 884, 866, 967, 787, 297, 219, 123, 661, 534, 748, …

rimozione > 750

297, 219, 123, 661, 534, 748, …

Per N e n "grandi", procedura dispendiosa: generazione di numeri *pseudo-casuali da Unif(0,1)*

Selezione sistematica – *campionamento* sistematico

Selezione dalla lista a intervalli regolari:

a partire da un inizio casuale, è selezionata una unità ogni *k*

$$k = \frac{N}{n}$$ = passo di campionamento

Es.: *N* = 2000, *n* = 250, k = 8

inizio: numero casuale $1 \le x \le 8$

unità nel campione con posizione: $x,\ x+k,\ x+2k,\ \ldots,\ x+(n-1)k$

se *x* = 5 *5, 13, 21…*

Se *k* non intero, preferibile considerare lista come circolare

Es.: *N* =1872, *n* = 250, *k = 7.488*

- si arrotonda ~ 7
- inizio casuale x da 1 a N (1-1872)
- si procede fino a ottenere 250 unità

(con solo arrotondamento –per eccesso/difetto- n variabile (es: n = 234 o n = 267)

Campionamento sistematico e proprietà CCS

1. probabilità di selezione costante $\pi_k = \frac{n}{N}$
2. probabilità di possibili insiemi di unità **variabile** **(campioni di unità separate da *k*. *Solo k campioni possibili*)**

esempio: campione con $k = 8$

P(1,2) = 0

P(1,9) = 1/8

ne consegue:

stimatore media campionaria non distorto ma derivazione varianza stimatore applicabile come CCS se:

ipotesi che lista sia ordinata più o meno casualmente (es. liste alfabetiche)

in realtà: problemi solo se la lista segue una qualche sequenza ciclica

Campionamento sistematico

Può essere usato quando la lista delle unità (frame) non è nota prima della selezione:

- lista "concettuale" costruita selezionando ogni *k-esima* unità fino ad *esaurire tutte le unità nella popolazione*
- dimensione *n* del campione nota solo alla fine della procedura di selezione

- per es., sapendo che un supermercato ha circa 1800 clienti al giorno (giornata di 10 ore continuate: 3 ogni minuto), volendo ottenere un campione rappresentativo di 200 clienti (k=1800/200=9) se ne intervista uno ogni 9, o, in alternativa, uno ogni 3 minuti.
- Deve essere possibilmente evitata ogni selezione diversa da quella predeterminata dal passo di campionamento (ad esempio quella in base alle caratteristiche delle persone).

Dimensione del campione casuale semplice /1

quanto grande deve essere *n*?

Varianza stimatore media campionaria $\text{var}(\bar{y}) = \frac{S^2}{n}(1-f)$

Nel caso di stima di una proporzione

$$\text{var}(\hat{p}) = \frac{p(1-p)}{n-1}\left(1-\frac{n}{N}\right) \approx \frac{p(1-p)}{n}$$

$\simeq n$ se n è grande

$\simeq 1$ se la popolazione è grande

fissato il grado di precisione desiderato con una certa probabilità, è possibile determinare il valore di *n* corrispondente

Dimensione del campione casuale semplice /2

quanto grande deve essere *n*?

Esempio: si richiede che la stima campionaria sia effettuata con una precisione del 2% con probabilità pari al 95%

Ovvero: intervallo di confidenza di $\hat{p}$ al livello 95%: $\hat{p} \pm 2\%$, $\Pr(\hat{p} \pm 2\%) = 95\%$

$$\hat{p} \pm \underbrace{1.96 SE(p)}_{2\%}$$

trascurando (*1-f*) e dividendo per *n*, $SE(p) \approx \sqrt{\frac{p(1-p)}{n}} \Rightarrow 0.02 = 1.96\sqrt{\frac{p(1-p)}{n}}$

Se si ipotizza *p nella popolazione* = 0.35 $\quad n = \frac{1.96^2\, p(1-p)}{0.0004} = 2185$

Se *N*=15.000 $\quad \Rightarrow n' = \frac{Nn}{N+n} = 1907$

se si tiene conto del tasso di risposta del 75% $\quad n'' = \frac{n'}{0.75} = 2543$

se precisione = 3% $\Rightarrow n = 971$ e $n' = 912$

se precisione = 1%: *n* = 10.084

Dimensione del campione casuale semplice /3

quanto grande deve essere *n*?

$$\left[\hat{p} - z\frac{s}{\sqrt{n}}; \hat{p} + z\frac{s}{\sqrt{n}}\right]$$

z ⇨ da $N(0,1)$

S ⇨ variabilità dei dati

n ⇨ numerosità del campione

- non c'è relazione con *N !!!*
- la precisione varia con $\sqrt{n}$

Dimensione del campione casuale semplice /4

quanto grande deve essere *n*?

Table 1-1 Precision of the sampling procedure used by the Gallup Poll as of 1972*

Population percentage	Sample size *n*						
	100	200	400	600	750	1000	1500
Near 10	7	5	4	3	3	2	2
Near 20	9	7	5	4	4	3	2
Near 30	10	8	6	4	4	4	3
Near 40	11	8	6	5	4	4	3
Near 50	11	8	6	5	4	4	3
Near 60	11	8	6	5	4	4	3
Near 70	10	8	6	4	4	4	3
Near 80	9	7	5	4	4	3	2
Near 90	7	5	4	3	3	2	2

SOURCE: George Gallup, *The Sophisticated Poll Watcher's Guide* (Princeton Opinion Press, 1972), p. 228.

* The table shows the range, plus or minus, within which the sample percentage p falls in 95% of all samples. This margin of error depends on the size of the sample and on the population percentage p. For example, when p is near 60%, 95% of all samples of size 1000 will have p between 56% and 64%, because the margin of error is ± 4%.

Vedere: https://www.surveymonkey.com/curiosity/how-many-people-do-i-need-to-take-my-survey/?ut_source=mp&ut_source2=random-sample-in-excel&ut_source3=inline

Alcune considerazioni a proposito di *n* /1

- livello di precisione

a. analisi per sottogruppi

↓

incroci di più variabili

es. tasso di disoccupazione per età, sesso, titolo di studio

la precisione dipende dal valore di n^* nel campione con quelle caratteristiche, non da n' totale

b. indagini multiscopo: più proporzioni e/o medie da stimare

che importanza (precisione) si deve dare? ⇨ diversa variabilità nella popolazione

scelta di variabile target e/o valutazione 'ragionata'

Alcune considerazioni a proposito di *n* /2

- stime di p o di S^2

  per proporzioni è più facile, perché p (1- p) non è molto variabile (p: {0.1;0.9})

  per S^2 è più complicato

  - indagini pilota
  - indagini precedenti
  - considerazioni su struttura della popolazione

Pesi (*weights*) di campionamento

π_k = probabilità di inclusione *unità k* nel campione

Possono essere usate per calcolare stime puntuali dei parametri di interesse

ω_k = *peso di campionamento* = $1/\pi_k$ = **Numero** di unità della popolazione *rappresentate dall'unità k*

CCS: $\pi_k = n/N$ *da cui* $\omega_k = 1/\pi_k = N/n$

CCS pesi tutti uguali: ogni unità nel campione *rappresenta se stessa e altre N/(n-1) unità (non selezionate) della popolazione (in totale N/n)*

$\Sigma\, \omega_k = \Sigma\, N/n = N$ $\Sigma\, \omega_k y_k = \Sigma (N/n)\, y_k = \hat{t}$

$\Sigma\, \omega_k y_k \,/\, \Sigma \omega_k = \hat{t}/\, N = \hat{y}$

Quando usare un CCS ?

- La popolazione è omogenea.
- Si dispone di buone liste dell'intera popolazione.
- Il costo per raggiungere ogni unità è omogeneo e non varia se si prevedesse l'uso di disegni più complessi.
- Si vogliono usare stimatori semplici.
- Si vogliono stimare relazioni complesse ed altri disegni hanno costi comparabili.