Esercizi campionamento per prova intermedia 12 dicembre 2024

1. Da una lista di 2000 scuole superiori con 1000 studenti ciascuna, è estratto un campione di 3000 studenti secondo una procedura a due stadi. Al primo stadio sono selezionate 100 scuole e al secondo sono estratti 30 studenti in ogni scuola selezionata. Il 30% degli studenti intervistati dichiara di possedere un computer personale e lo standard error di tale stima è pari a 1.4%. Ignorando il fattore di correzione per popolazione finita e con l'approssimazione $n - 1 = n$, stimare:

   a. il *deff* per la percentuale campionaria;
   stimadeff=(.000196/.00007)=2.8

   b. il coefficiente di correlazione intraclasse ρ entro le scuole per il possesso di computer da parte degli studenti;
   stimarho = (*stimadeff* − 1)/(l − 1) = .0621, con l = dimensione campione studenti entro unità stadio 1. (commento: livello di omogeneità molto modesto)

   c. lo standard error della percentuale campionaria nel caso di un campione che prevede la selezione di 300 scuole e di 10 studenti per scuola;
   $stimadeff_{new}$ = 1+9 *.0621= 1.558, $stimase_{new}$= .0104 (commento: di poco inferiore al precedente poichè valore *stimarho* molto modesto)

   d. i campioni così selezionati sono auto-ponderanti? Motivare la risposta.
   Sì, dato che le unità di stadio 1 hanno stessa dimensione ed è selezionato lo stesso numero di unità di stadio 2 entro le unità di stadio 1 estratte.

2. In un CCS di 1000 unità selezionate da 10000 ricoverati di un ospedale, 35 soggetti sono affetti da una malformazione cardiaca. La tabella riporta i risultati campionari secondo l'età dei soggetti intervistati:

Età	N.ro soggetti	N.ro affetti da malformazione cardiaca	%
<= 25	100	25	25
26-39	500	7	1.4
>= 40	400	3	0.75
Totale	1000	35	3.5

   a. calcolare l'errore standard della stima della proporzione di affetti da malformazione;
   stimase = .0055

   b. sulla base delle percentuali in ultima colonna, la stratificazione avrebbe potuto fornire una stima migliore? Motivare la risposta.
   nonostante si tratti di dati solo campionari, la differenza nelle % fa pensare che analoga differenza si verifichi a livello di popolazione. Una procedura di stratificazione potrebbe ridurre la varianza della stima rispetto al CCS.

   c. supponendo che ci siano 2000 soggetti sino a 25 anni, 6000 tra 26-39 anni e oltre 2000 di 40 anni, nella situazione ipotizzata al punto [b.] stimare la proporzione di affetti da malformazione e la sua varianza.
   $stimap_{new}$ = .0599; $stimavar_{new}$ = .000082 (attenzione: risulta maggiore di quella al punto 1 anche se stratificato, poichè varianza stimata strato 1 molto elevata ma strato con frazione camp.to più bassa.

d. supponendo di aver estratto un campione stratificato di 333 unità per ciascuna classe di età e che il numero di affetti da malformazione sia quello in tabella con numerosità dei gruppi come al punto [b.], stimare la proporzione di affetti da malformazione e la sua varianza.

new% nei 3 strati: 7.5, 2.1, .9., $stimap_{newnew} = .0294$, $stimavar_{newnew} = .000031$ (commento: nuova stima della varianza molto inferiore al punto [b.] ma molto simile a punto [a.])

3. Una azienda vuole stimare il n.ro di giorni persi per malattia dai suoi impiegati distribuiti in 8 divisioni, ciascuna con un n.ro diverso di impiegati, come riportato nella tabella seguente:

Divisione	1	2	3	4	5	6	7	8
Impiegati	1200	450	2100	860	2840	1910	390	3200

a. mostrare come selezionare un campione di 3 divisioni nelle quali rilevare il numero di giorni persi, tenendo conto in modo opportuno della diversa numersità delle divisioni.

descrivere metodo della cumulata (più adatto) o metodo di Lahiri.

b. si supponga siano state estratte le divisioni 3, 6 e 8, per le quali i giorni persi per malattia sono pari rispettivamente a 4320, 4160 e 5790. Stimare il numero medio di giorni persi per malattia da ogni impiegato dell'azienda;

dato che la selezione delle 3 divisioni è PPS, il numero medio di giorni persi per malattia si ottiene come media semplice delle 3 medie di ciascuna divisione selezionata (pari a 2.057, 2.178 e 1.809), $stimamedia = 2.015$

c. valutare la varianza dello stimatore.

analogamente, *stimavar* si ottiene dividendo la devianza delle 3 medie dalla media generale per $m(m-1)$, con $m =$ numero di divisioni estratte. $stimavar = .012$

4. Una azienda intende condurre una indagine campionaria di *customer satisfaction* sui suoi 10000 clienti. Il questionario contiene items misurati su scala da 0 (=estremamente non soddisfatto) a 100 (=estremamente soddisfatto) che sono usati per calcolare un indice di soddisfazione. Prima di condurre l'indagine effettiva, è effettuata una pilota su un CCS di 20 clienti. I valori dell'indice di soddisfazione sono:

100	88	72	81	80	69	84	83	65	69	90	65	70	80	90	74	70	96	62	67

a. calcolare la media campionaria dell'indice di soddisfazione e lo standard error.

$stimamedia = 77.75$, $stimase = 2.476$

b. costruire l'intervallo di confidenza al 95% per l'indice medio di soddisfazione dei 10000 clienti. L'ufficio qualità dell'azienda intende raggiungere un target pari a 80. Che considerazioni si possono fare al riguardo?

[73; 83] contiene il valore 80

c. l'azienda decide di avviare l'indagine, con la richiesta di un margine di errore per l'intervallo di confidenza al 95% non superiore a 2. Calcolare la corrispondente dimensione campionaria sulla base dei risultati dell'indagine pilota.

$n = 118$

5. Un gruppo di 124 nazioni è stato suddiviso in 4 strati sulla base della popolazione di tre anni prima. Nei primi 3 strati sono stati estratti campioni indipendenti, rispettivamente di dimensione 6, 6, 11, mentre nell'ultimo sono state selezionate tutte le 7 nazioni dello strato, per una dimensione totale del campione stratificato $n = 30$. I dati osservati (in milioni) sono riportati in tabella:

Strato	Pop.ne 3 anni prima	N_h	n_h	$\sum_i y_i$	$\sum_i y_i^2$
1	0-9.99	70	6	31.1	234.95
2	10-29.99	29	6	119.6	2613.26
3	30-99.99	18	11	621.0	40122.36
4	100 e oltre	7	7	2671.4	1769052.90

a. determinare rispettivamente la stima del totale di popolazione e la varianza in ogni strato e nelle 124 nazioni;

stimatotalistr = 362.833, 578.067, 1016.182, 2671.4, *stimatotale* = 4628.482 (somma delle stime dei 4 totali di strato)

stimavartotstr = 11016.32, 5098.437, 5800.032, 0, *stimavartot* = 21914.789 (somma delle stime delle 4 varianze dei totali di strato)

b. costruire un intervallo di confidenza al 95% per il totale di popolazione delle 124 nazioni e commentare il risultato;

con riferimento a distribuzione Normale: [4338.331; 4918.633]

c. quale sarebbero state le dimensioni dei campioni per strato nel caso di una allocazione proporzionale del campione di nazioni?

frazione campionamento = $30/124 = .24$, $n_1 = 17$, $n_2 = 7$, $n_3 = 4$, $n_4 = 2$

d. per quale motivo a vostro parere non è stata utilizzata l'allocazione proporzionale del campione di nazioni?

oltre metà campione finale sarebbe formato da nazioni "piccole'.