Cognome **Nome**

Matricola

Tecniche di indagine statistica - Prova 2

19 dicembre 2023

1. Quesiti V/F

- Poichè non è noto chi risponderà, il campionamento su base volontaria può essere assimilato ad un campionamento probabilistico F ☐
- Quando la frazione di campionamento n/N è molto bassa il campione non è rappresentativo F ☐
- In un campione probabilistico, ogni n-pla di unità ha una probalità strettamente maggiore di zero di essere estratta F ☐
- Il campionamento per quote è analogo al campionamento stratificato F ☐
- La numerosità campionaria non influenza la precisione delle stime F ☐
- Il coefficiente di correlazione intraclasse è sempre positivo F ☐
- Con un campione probabilistico si evita ogni forma di distorsione da selezione F ☐
- La selezione sistematica produce un campione con le stesse caratteristiche di un CCS F ☐
- S^2 può essere calcolata moltiplicando σ^2 per $(N-1)/N$ F ☐
- Il campione a grappoli equivale ad un campione CCS di totali della variabile di interesse V ☐
- In uno strato molto variabile rispetto alla variabile di interesse conviene selezionare più unità V ☐
- In un campione a 2 stadi auto-ponderante si deve tener conto del numero di unità di primo stadio F ☐
- Il metodo di Lahiri è usato per la selezione di unità di primo stadio V ☐
- Nella post-stratificazione si usano i pesi di campionamento F ☐
- Se le non risposte sono di tipo MAR, possono essere ignorate nella stima dei parametri della variabile Y F ☐

2. La tabella seguente riporta i dati (fittizi) relativi a due popolazioni, ciascuna composta da tre (3) gruppi (unità di primo stadio). Ogni gruppo contiene a sua volta tre (3) unità sulle quali sono stati osservati i valori relativi alla medesima variabile Y.

	Popolazione A			Popolazione B		
Gruppo 1	10	20	30	9	10	11
Gruppo 2	11	20	32	17	20	20
Gruppo 3	9	17	31	31	32	30

   a. calcolare la media e la varianza di Y nelle due popolazioni e per ciascun gruppo. Commentare i risultati.
   $M(Y)_A = M(Y)_B = 20$ $S^2_A = S^2_B = 84.5$

	Popolazione A		Popolazione B	
	$M(Y)$	S^2	$M(Y)$	S^2
Gruppo 1	20	100	10	1
Gruppo 2	21	111	19	3
Gruppo 3	19	124	31	1

Pop_A: medie di gruppo molto simili, variabilità elevata *entro* i gruppi (devianza = 670)
Pop_B: medie di gruppo diverse, variabilità elevata *tra* i gruppi (devianza = 666)

b. sapendo che ICC_A = -.4867 e ICC_B = .9803 che considerazioni si possono trarre ai fini della scelta del tipo di campionamento probabilistico da adottare nelle due popolazioni?

Pop_A: maggiore variazione avviene entro i gruppi e variazione molto contenuta tra i gruppi. ICC negativo riflette tale situazione: unità nello stesso gruppo sono meno *simili* che unità prese casualmente dall'intera popolazione. Il campionamento a grappolo sarebbe più efficiente di un CCS.
Pop_A: maggiore variazione avviene tra i gruppi e variazione molto contenuta entro i gruppi. ICC positivo e molto vicino a 1 indica che non si otterebbe molta nuova informazione selezionando più di una unità per gruppo. Il campionamento a grappolo sarebbe molto meno efficiente di un CCS mentre il campionamento stratificato (se attuabile) con strati dati dai gruppi potrebbe essere preferibile.

3. Nella preparazione di una indagine sulle famiglie del Friuli Venezia Giulia, il campione pilota è stato suddiviso casualmente in tre gruppi ai quali sono stati proposti, rispettivamente, nessun incentivo (gruppo A), 20 euro (gruppo B) e 35 euro (gruppo C) per la partecipazione all'indagine. Alcuni dati relativi ai costi d'indagine (per famiglia intervistata) e al tasso di risposta nei tre gruppi sono riportati nella tabella seguente:

	Gruppo A	Gruppo B	Gruppo C
Compenso intervistatore	125	95	90
Costo incentivo	0	20	35
Costo totale	125	115	125
Tasso di risposta (%)	64	71	74

a. perchè il campione è stato suddiviso casualmente in tre gruppi? Interpretare i risultati;
Disporre di risultati "sperimentali" per valutare l'effetto dell'incentivo sul tasso di risposta. L'incentivo ha un effetto positivo sul tasso di risposta.

b. se il campione pilota fosse un campione casuale semplice delle famiglie residenti in regione, approssimativamente e senza ricorrere a calcoli quale dovrebbe essere la sua dimensione per garantire che la stima del tasso di risposta nel gruppo A sia significativamente inferiore a quella dei gruppi B e C ?
Meglio se intorno a 1000 e comunque non inferiore a 700.

4. Una azienda vuole stimare il n.ro medio di giorni di assenza all'anno dei suo 300 dipendenti. Per procedere alla selezione del campione, sono disponibili due diverse liste dei dipendenti:

- Lista A, dipendenti nei 4 reparti dell'azienda, rispettivamente con numerosità pari a 75, 60, 100, 65;
- Lista B, dipendenti per 3 classi di età, rispettivamente con numerosità pari a 35, 200, 65.

Si supponga, inoltre, che siano disponibili le seguenti informazioni sulla variabilità dei giorni di assenza per reparto e per classe di età:

Lista A	$s_1^2 = 12$	$s_2^2 = 21$	$s_3^2 = 46$	$s_4^2 = 35$
Lista B	$s_1^2 = 35$	$s_2^2 = 40$	$s_3^2 = 23$	

a. con un campione di $n = 20$ dipendenti, quanti dipendenti dovrebbero essere selezionati (con procedura CCS) da ogni strato per massimizzare l'efficienza dello stimatore della media usando la Lista A? E nella Lista B?

Lista A

Strato	s_h	N_h	$N_h s_h$
1	3.464	75	259.80
2	4.583	60	1274.98
3	6.782	100	678.20
4	5.916	65	384.55
Totale		300	1597.53

$n_1 = 3.3[3], n_2 = 3.44[3], n_3 = 8.5[9], n_4 = 4.8[5]$

Lista B: $n_1 = 2.3[2], n_2 = 14.18[14], n_3 = 3.5[4]$

b. sulla base delle numerosità campionarie per strato fatta al punto (a), quale tra le due stratificazioni fornisce lo stimatore più efficiente?

Lista A: $V(\hat{y}) = 1.326$

Lista B: $V(\hat{y}) = 1.659$

c. rispondere al punto (a) nell'ipotesi che non siano note le informazioni sulla variabilità dei giorni di assenza seconde le due liste;

stratificato proporzionale

Lista A: $n_1 = 5, n_2 = 4, n_3 = 7, n_4 = 4$

Lista B: $n_1 = 2, n_2 = 14, n_3 = 4$

d. confrontare e interpretare i risultati ottenuti ai punti precedenti.

Risultati allocazione ottimale e proporzionali sono diversi con Lista A mentre sono gli stessi con Lista B.

5. In tre città diverse, Trieste, Bologna e Roma si estrae un campione casuale semplice senza reinserimento di persone pari all'1% della popolazione maggiorenne delle tre città per stimare la proporzione di favorevoli ai pagamenti mediante POS. In quale città la varianza campionaria della stima è più bassa ? Motivare la risposta.

Poichè $V(\hat{p}) = [(N-n)/(N-1)][p(1-p)/n]$, la varianza campionaria della stima è più bassa nella città con il campione di numerosità maggiore, ovvero Roma.