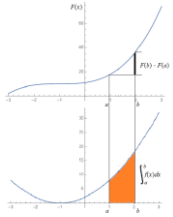
ISSN: 2037-6367

MatematicaMente

Pubblicazione mensile della sezione veronese della MATHESIS – Società Italiana di Scienze Matematiche e Fisiche – Fondata nel 1895 – Autorizzazione del Tribunale di Verona n. 1360 del 15 – 03 – 1999 – I diritti d'autore sono riservati. Direttore: Luciano Corso – Redazione: Alberto Burato, Fabrizio Giugni, Michele Picotti, Sisto Baldo – Via IV Novembre, 11/b – 37126 Verona – tel e fax (045) 8344785 – 338 6416432 – e-mail: lcorso@iol.it – info@mathesisverona.it – Stampa in proprio – Numero 223 – Pubblicato il 06 – 04 – 2017



Campionamento stratificato ottimale

di Aleksandr A. Čuprov e Jerzy Neyman

di Luciano Corso

Consideriamo una popolazione di cardinalità N di una variabile statistica X con media aritmetica μ e varianza σ^2. Dividiamo questa popolazione in k strati (uno strato è un gruppo di dati selezionato in coerenza alla ricerca di un dato carattere sensibile alla proprietà posseduta dallo strato):

Strato 1	Strato 2	...	Strato k	Popolazione
N_1	N_2	...	N_k	N
μ_1	μ_2	...	μ_k	μ
σ_1^2	σ_2^2	...	σ_k^2	σ^2
n_1	n_2	...	n_k	n
$\bar{X}_1$	$\bar{X}_2$	...	$\bar{X}_k$	$\bar{X}$
$s^2_{\bar{X}_1}$	$s^2_{\bar{X}_2}$	...	$s^2_{\bar{X}_k}$	$s^2_{\bar{X}}$

Estraiamo quindi un campione di n elementi dalla popolazione in modo tale che, per ogni strato, siano rispettivamente n_1, n_2, …, n_k gli elementi estratti. C'è quindi un vincolo all'estrazione campionaria:

$$\sum_{i=1}^{k} n_i = n \,. \tag{1}$$

All'interno di ogni strato i, la media aritmetica campionaria è:

$$\bar{X}_i = \frac{1}{n_i} \cdot \sum_{j=1}^{n_i} x_{i,j} \qquad \forall i = 1, ..., k \,. \tag{2}$$

La media aritmetica campionaria totale degli strati è data da:

$$\bar{X} = \sum_{i=1}^{k} \frac{N_i}{N} \cdot \bar{X}_i \,, \tag{3}$$

dove N_i / N è il peso relativo di ogni media di strato (media aritmetica ponderata).
(3) è uno stimatore corretto di μ, infatti

$$M(\bar{X}) = \sum_{i=1}^{k} \frac{N_i}{N} M(\bar{X}_i) = \sum_{i=1}^{k} \frac{N_i}{N} \mu_i = \mu \,. \tag{4}$$

Se le estrazioni campionarie negli strati sono indipendenti si ha poi che:

$$Var(\bar{X}) = \sigma^2_{\bar{X}} = \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \cdot Var(\bar{X}_i) = \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \cdot \frac{\sigma_i^2}{n_i} \,. \tag{5}$$

Se, invece le estrazioni campionarie negli strati sono in blocco si ottiene

$$\begin{aligned} Var(\bar{X}) = \sigma^2_{\bar{X}} &= \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \cdot Var(\bar{X}_i) \\ &= \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \cdot \frac{N_i - n_i}{N_i - 1} \cdot \frac{\sigma_i^2}{n_i} \,. \end{aligned} \tag{6}$$

Per la verifica di (6) e, in particolare, per giustificare la presenza del così detto fattore di correzione: $(N_i - n_i) / (N_i - 1)$ si rinvia a [B.2]. A questo punto si distingue tra campionamento bernoulliano (prove indipendenti) e campionamento in blocco.

1. Campionamento bernoulliano negli strati

Nel campionamento bernoulliano, la varianza della media aritmetica campionaria totale degli strati è quindi data da (5); cioè:

$$\sigma^2_{\bar{X}} = \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \cdot \frac{\sigma_i^2}{n_i} \,. \tag{7}$$

dove s al pedice viene aggiunto per indicare che si tratta di un campionamento stratificato. A questo punto c'è da trovare il minimo di questa funzione con il vincolo (1):

$$\begin{cases} Min[\sigma^2_{\bar{X}}] \\ \sum_{i=1}^{k} n_i = n \end{cases} \tag{8}$$

I minimi vincolati sono piuttosto complessi da cercare. Esiste, però, un teorema di Lagrange che afferma che se una funzione ammette un minimo vincolato, allora questo può essere calcolato da una funzione che risulta dalla combinazione lineare della funzione data con il prodotto di un opportuno parametro con il vincolo dato. Cioè:

$$L(n_1, n_2, ..., n_k, \lambda) = \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \cdot \frac{\sigma_i^2}{n_i} + \lambda \cdot \left(\sum_{i=1}^{k} n_i - n\right) , \tag{9}$$

dove λ è il parametro da stimare e la ricerca del miglior campionamento è sulle variabili n_1, n_2, …, n_k. Si deve procedere al calcolo delle derivate parziali prime rispetto alle $k + 1$ variabili e risolvere il sistema ponendo tali derivate uguali a zero (non occorre studiare il segno delle derivate seconde in quanto esse generano una matrice semi definita positiva e quindi là dove si annullano le derivate parziali prime si è in presenza di un minimo). Si ha:

$$\begin{cases} \dfrac{\partial}{\partial n_i} L(n_1, n_2, ..., n_k; \lambda) = -\dfrac{\left(\frac{N_i}{N}\right)^2 \cdot \sigma_i^2}{n_i^2} + \lambda = 0 \,, \quad i = 1, 2, ..., k \\ \dfrac{\partial}{\partial \lambda} L(n_1, n_2, ..., n_k; \lambda) = \sum_{i=1}^{k} n_i - n = 0 \end{cases} \tag{10}$$

Dalla prima e dalla seconda equazione si ottiene:

$$n_i = \frac{\frac{N_i}{N} \cdot \sigma_i}{\sqrt{\lambda}} \quad , \quad \sum_{i=1}^{k} n_i = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{k} \frac{N_i}{N} \cdot \sigma_i \quad , \quad \sqrt{\lambda} = \frac{\sum_{i=1}^{k} \frac{N_i}{N} \cdot \sigma_i}{n}$$

e quindi:

$$n_i = \frac{n \cdot \frac{N_i}{N} \cdot \sigma_i}{\sum_{j=1}^{k} \frac{N_j}{N} \cdot \sigma_j} = \frac{n \cdot N_i \cdot \sigma_i}{\sum_{j=1}^{k} N_j \cdot \sigma_j} \,, \quad \forall i \,. \tag{11}$$

2. Campionamento in blocco negli strati

Nel caso di campionamento in blocco (estrazioni senza riposizione) si procede nello stesso modo, ma a partire da (6). Si ha:

$$\sigma^2_{\bar{X}} = \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \cdot \frac{\sigma_i^2}{n_i} \cdot \frac{N_i - n_i}{N_i - 1} \,. \tag{12}$$

Da (8) in questo caso la funzione di Lagrange è:

$$L(n_1, n_2, ..., n_k, \lambda) = \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \cdot \frac{\sigma_i^2}{n_i} \cdot \frac{N_i - n_i}{N_i - 1} + \lambda \cdot \left(\sum_{i=1}^{k} n_i - n\right) \tag{13}$$

e quindi

$$\begin{cases} \dfrac{\partial}{\partial n_i} L\left(n_1, n_2, ..., n_k; \lambda\right) = 0\,, \quad i = 1, 2, ..., k \\ \dfrac{\partial}{\partial \lambda} L\left(n_1, n_2, ..., n_k; \lambda\right) = 0 \end{cases} \tag{14}$$

$$\begin{cases} -\dfrac{\left(\frac{N_i}{N}\right)^2 \cdot \sigma_i^2}{n_i^2} \cdot \left(\dfrac{N_i - n_i}{N_i - 1}\right) + \dfrac{\left(\frac{N_i}{N}\right)^2 \cdot \sigma_i^2}{n_i^2} \cdot \left(-\dfrac{1}{N_i - 1}\right) + \lambda = 0 \quad , \qquad i = 1, 2, ..., k \\ \sum_{i=1}^{k} n_i - n = 0 \end{cases}$$

Risolvendo questo sistema si ottiene:

$$\sqrt{\lambda} = \frac{\displaystyle\sum_{i=1}^{k} \frac{N_i}{N} \cdot \sigma_i \cdot \sqrt{\frac{N_i}{N_i - 1}}}{n} \quad , \quad n_i = \frac{n \cdot \dfrac{N_i}{N} \cdot \sigma_i \cdot \sqrt{\dfrac{N_i}{N_i - 1}}}{\displaystyle\sum_{j=1}^{k} \frac{N_j}{N} \cdot \sigma_j \cdot \sqrt{\frac{N_j}{N_j - 1}}}\,. \tag{15}$$

Per chiarire meglio l'argomento, facciamo un esempio: Le altezze in centimetri di N =700 individui, raggruppati in classi di frequenza, sono: x_1 = 180 con frequenza N_1 = 150 e σ_1 = 2, x_2 = 168 con N_2 = 70 e σ_2 = 3, x_3 = 171 con N_3 = 50 e σ_3 = 4, x_4 = 176 con N_4 = 200 e σ_4 = 2,5, x_5 = 170 con N_5 = 100 e σ_5 = 1,5, x_6 = 182 con N_6 = 80 e σ_6 = 1, x_7 = 165 con N_7 = 50 e σ_7 = 1,8, dove x_i è il valore centrale della classe.
Si presenti empiricamente il guadagno di una stratificazione per classi di altezza secondo Neyman rispetto a un campionamento casuale stratificato proporzionale ($N_i : n_i = N : n$) ponendo come vincolo un campione di n = 50 unità. Con il metodo di Neyman e campionamento bernoulliano (si applica la (11) e si arrotonda opportunamente) si ottiene:
n_1 = 10, n_2 = 7, n_3 = 6, n_4 = 16, n_5 = 5, n_6 = 3, n_7 = 3
con deviazione standard pari a

$Neyman$: $\sigma_{\bar{x}} \cong 0{,}310366$.

Con il metodo della stratificazione proporzionale si ottiene
n_1 = 11, n_2 = 5, n_3 = 3, n_4 = 14, n_5 = 7, n_6 = 6, n_7 = 4
con deviazione standard pari a

$Proporzionale$: $\sigma_{\bar{x}} \cong 0{,}558639$,

che è un valore quasi doppio del precedente.
Si noti, quindi, il vantaggio di precisione che si ottiene con il metodo di Čuprov-Neyman.

Riferimenti bibliografici: [B.1] Riporto solo un libro di testo, quello in dotazione all'ITIS G. Marconi di Verona (in Matematica Applicata, corso d'Informatica) prima della riforma Gelmini: Gambotto Manzone A. M. & Susara Longo C., *Inferenza statistica e Ricerca Operativa*, pagg. 80-91, Ed. Tramontana, anno 2007. Alle pagine 86 e 87 si presenta il problema della ripartizione ottimale di Neyman lasciando agli studenti "per esercizio" il compito di dimostrare i risultati del metodo sopra esposti. [B.2] Guerrini A. & Corso L., *Varianza di medie aritmetiche campionarie di campioni finiti estratti in blocco da popolazioni finite*, MatematicaMente n. 110, Mathesis sez. VR, anno 2006, ISSN: 2037-6367.

Campionamento: alcune note, in pillole

(*di Luciano Corso*) Come è noto, quando si vogliono studiare i caratteri statistici di una popolazione di dati e non si può, per varie ragioni, estendere l'indagine all'universo dei dati possibili, in statistica si conviene di orientare l'investigazione su un campione di questi dati. Un campione è un gruppo di dati estratti da una popolazione secondo modalità convenute. I metodi che maggiormente vengono usati per selezionare le unità campionarie dovrebbero sempre tener conto che alla fine il campione estratto deve essere in qualche modo rappresentativo delle caratteristiche statistiche presenti nella popolazione. La rappresentatività dei dati campionari è una proprietà molto gradita, ma sappiamo che non è di facile ottenimento. Vediamo che cosa può succedere se da una popolazione finita si estraggono campioni finiti e analizziamo con un esempio molto elementare le inevitabili distorsioni associate al campionamento.

Consideriamo una popolazione costituita da N = 6 elementi e siano essi i primi sei numeri naturali: X = {1, 2, 3, 4, 5, 6}. Vediamo già molto della struttura statistica di questa popolazione e possiamo anche calcolare le sue "statistiche" più significative. In particolare, si determinano facilmente: la media aritmetica, μ = 3,5; la varianza, $\sigma^2 \cong 2{,}916$; la distribuzione di probabilità: $h(x)$ = p =1 / $|X|$ con $x \in X$ (uniforme discreta).
Normalmente, però, la struttura statistica di una popolazione di dati non è conosciuta ed è necessario, in questi casi, lavorare su un campione di osservazioni.

Per selezionare unità campionarie si possono usare vari metodi a seconda del tipo di indagine che si vuole fare. Nel campionamento casuale semplice, i tipi di estrazioni possono essere riassunte nella tabella 1.

Da una popolazione di N unità si estrae un campione di n unità; in generale, $n \le N$ ma non necessariamente; possono esserci campioni i cui dati sono più numerosi di quelli della popolazione di provenienza. Le estrazioni possono essere di quattro tipi a seconda che ci sia riposizione (reinserimento) delle unità estratte e conti l'ordine di selezione.

Tabella 1

		Ordine: sì	Ordine: no
Riposizione	sì	N^n	$\binom{N+n-1}{n}$
	no	$N^{\lceil n \rceil}$	$\binom{N}{n}$

Nel campionamento cosiddetto bernoulliano c'è riposizione e l'ordine conta. Nel campionamento in blocco non c'è riposizione. Per capire ciò che succede nei 4 casi esposti in tabella 1, consideriamo un campione di n = 2 unità e presentiamo i possibili campioni ottenibili a partire da X:

Caso (1,1)

11	12	13	14	15	16
21	22	23	24	25	26
31	32	33	34	35	36
41	42	43	44	45	46
51	52	53	54	55	56
61	62	63	64	65	66

Caso (1,2)

11	12	13	14	15	16
	22	23	24	25	26
		33	34	35	36
			44	45	46
				55	56
					66

Caso (2,1)

	12	13	14	15	16
21		23	24	25	26
31	32		34	35	36
41	42	43		45	46
51	52	53	54		56
61	62	63	64	65	

Caso (2,2)

12	13	14	15	16	
	23	24	25	26	
		34	35	36	
			45	46	
				56	

.

Se dovessimo estrarre un campione di n = 6 unità, solo nei casi (2,1) e (2,2) si avrebbe a disposizione un'informazione completa dello stato della popolazione, semplicemente perché il campione così sarebbe costituito da tutti i dati della popolazione. Negli altri casi una distorsione si manifesterebbe comunque. Vediamo che succede, per esempio, nel caso di un campione di n=2 unità e sia esso (x_1= 2, x_2 =4). Calcoliamo media aritmetica e varianza campionarie: $\bar{x} = 3$ e $s^2 = 1$ (Ricordo che la varianza è il quadrato della distanza media euclidea dei dati dalla media aritmetica del gruppo). Mi accorgo che sia $\bar{x}$ sia s^2 differiscono da μ e σ^2 della popolazione. Se facciamo il calcolo di tutte medie aritmetiche campionarie dei possibili campioni che potremmo avere, scopriamo che $M(\bar{x}) = \mu$, dove l'applicazione M è la media aritmetica. Se, invece, calcolassimo $M(s^2)$ che succederebbe? Lasciamo ai lettori il compito di rispondere. Quando uno stimatore ha la proprietà di dare stime di un parametro incognito di una popolazione di dati centrate su di esso, si dice che lo stimatore è corretto. Nel nostro caso $\bar{x}$ è uno stimatore corretto di μ.

Riferimenti bibliografici: Piccolo Domenico, *Statistica*, il Mulino, Bologna, 1998. Possono, comunque essere usati molti altri manuali di statistica inferente presenti sul mercato.