

Popolazione e campione

- *Popolazione*: totalità dei casi (unità statistiche) sede del fenomeno oggetto di studio.

- *Rilevazione censuaria e campionaria*

Data una popolazione finita un **campione casuale** di ampiezza n si ottiene estraendo a sorte **con reimmissione** n unità dalla popolazione

- *Popolazione finita e infinita* (astrazione utile a fini didattici)

$f(x)$ = modello descrittivo della popolazione

- *Campione casuale di ampiezza n* : variabile aleatoria multipla (X_1, X_2, \dots, X_n)

le cui componenti sono **indipendenti e identicamente distribuite** con distribuzione $f(x)$, dove $f(x)$ denota il modello descrittivo della popolazione. Indichiamo con (x_1, x_2, \dots, x_n) il campione effettivamente osservato: un punto nello spazio euclideo ad n dimensioni

- *Spazio campionario* (Ω) : insieme di tutti i possibili campioni estraibili dalla popolazione.

Spazio campionario discreto o continuo a seconda se X è discreta o continua

- *Distribuzione del campione*:

$$f(x_1, x_2, \dots, x_n) = f(x_1)f(x_2) \cdots f(x_n)$$

questa è una funzione di **probabilità** (caso discreto) o di **densità** (caso continuo) a seconda della natura di $f(x)$ (o, equivalentemente, di X)

INDIPENDENZA delle osservazioni campionarie

ESEMPIO:

Siamo interessati a studiare la composizione delle famiglie di una certa collettività.

Per semplicità, immaginiamo che vi siano solo 8 famiglie, che etichettiamo con

1, 2, ..., 8, e che il numero di componenti per ciascuna di esse sia dato da

$$\begin{cases} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & \text{etichetta della famiglia} \\ 2 & 5 & 4 & 3 & 3 & 4 & 4 & 4 & \text{numero di componenti} \end{cases}$$

Il numero di componenti nella popolazione è descritto da

$$X = \begin{cases} 2 & 3 & 4 & 5 & \text{numero componenti} \\ 1/8 & 2/8 & 4/8 & 1/8 & \text{proporzione di famiglie} \end{cases}$$

Il numero medio di componenti nella popolazione è $E(X) = 3.625$ e la varianza nella popolazione è $V(X) = 0.734$.

Supponiamo ora di estrarre un campione casuale (cioè con reimmissione) di ampiezza due dalla popolazione descritta

Lo spazio dei possibili *dati campionari*, ossia di tutte le coppie di valori del carattere "numero di componenti" che potremmo rilevare (spazio campionario) è rappresentato nella tabella dove indichiamo in ogni casella, sia le possibili coppie di famiglie che potremmo estrarre, sia (in grassetto) il corrispondente numero di componenti.

I/II	1	2	3	4	5	6	7	8
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)	(1, 7)	(1, 8)
	(2, 2)	(2, 5)	(2, 4)	(2, 3)	(2, 3)	(2, 4)	(2, 4)	(2, 4)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)	(2, 7)	(2, 8)
	(5, 2)	(5, 5)	(5, 4)	(5, 3)	(5, 3)	(5, 4)	(5, 4)	(5, 4)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)	(3, 7)	(3, 8)
	(4, 2)	(4, 5)	(4, 4)	(4, 3)	(4, 3)	(4, 4)	(4, 4)	(4, 4)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)	(4, 7)	(4, 8)
	(3, 2)	(3, 5)	(3, 4)	(3, 3)	(3, 3)	(3, 4)	(3, 4)	(3, 4)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)	(5, 7)	(5, 8)
	(3, 2)	(3, 5)	(3, 4)	(3, 3)	(3, 3)	(3, 4)	(3, 4)	(3, 4)
6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)	(6, 7)	(6, 8)
	(4, 2)	(4, 5)	(4, 4)	(4, 3)	(4, 3)	(4, 4)	(4, 4)	(4, 4)
7	(7, 1)	(7, 2)	(7, 3)	(7, 4)	(7, 5)	(7, 6)	(7, 7)	(7, 8)
	(4, 2)	(4, 5)	(4, 4)	(4, 3)	(4, 3)	(4, 4)	(4, 4)	(4, 4)
8	(8, 1)	(8, 2)	(8, 3)	(8, 4)	(8, 5)	(8, 6)	(8, 7)	(8, 8)
	(4, 2)	(4, 5)	(4, 4)	(4, 3)	(4, 3)	(4, 4)	(4, 4)	(4, 4)

La distribuzione congiunta del campione (X_1, X_2) può essere ottenuta facilmente dalla tabella precedente, ed è descritta nella seguente tabella a doppia entrata dove, in ciascuna casella riportiamo anche, in grassetto, il corrispondente valore della media campionaria:

X_1/X_2	2	3	4	5	$p_{X_1}(\cdot)$
2	1/64 2	2/64 2.5	4/64 3	1/64 3.5	1/8
3	2/64 2.5	4/64 3	8/64 3.5	2/64 4	2/8
4	4/64 3	8/64 3.5	16/64 4	4/64 4.5	4/8
5	1/64 3.5	2/64 4	4/64 4.5	1/64 5	1/8
$p_{X_2}(\cdot)$	1/8	2/8	4/8	1/8	

Notiamo che X_1 e X_2 sono indipendenti e identicamente distribuiti, con la stessa distribuzione della popolazione X .

La distribuzione della media campionaria $\bar{X} = \frac{X_1+X_2}{2}$ si ottiene dalla tabella precedente, e risulta

$$\bar{X} = \begin{cases} 2 & 2.5 & 3 & 3.5 & 4 & 4.5 & 5 \\ \frac{1}{64} & \frac{4}{64} & \frac{12}{64} & \frac{18}{64} & \frac{20}{64} & \frac{8}{64} & \frac{1}{64} \end{cases}$$

Il valore atteso di \bar{X} è

$$E(\bar{X}) = 2\frac{1}{64} + 2.5\frac{4}{64} + \dots + 5\frac{1}{64} = 3.625 .$$

La varianza di \bar{X} è

$$V(\bar{X}) = (2^2\frac{1}{64} + 2.5^2\frac{4}{64} + \dots + 5^2\frac{1}{64}) - 3.625^2 = 0.3672$$

Si può verificare che $E(\bar{X}) = E(X)$ e $V(\bar{X}) = \frac{V(X)}{2}$.

- *Parametro* = ω = costante caratteristica della popolazione

Nell'esempio precedente il parametro di interesse è il numero di componenti della famiglia

- *Parametri scalari e vettoriali*. Per esempio:

Popolazione Bernoulliana: il parametro è p (scalare)

Popolazione Normale: il parametro è $\omega = (\mu, \sigma^2)$ (vettoriale)

- *Spazio dei parametri* = Ω = insieme di tutti i valori plausibili per il parametro.

Per esempio:

Popolazione Normale: $\Omega = (\mathcal{R}, \mathcal{R}^+)$

- *Inferenza statistica:* risalire dai dati di un campione alle caratteristiche rilevanti della popolazione o meglio di un parametro della popolazione (assunto un modello teorico per la popolazione stessa)
- *Stima dei parametri* (puntale o per intervallo): sulla base del campione assegno al parametro di interesse un valore o un insieme di valori
- *Verifica delle ipotesi:* faccio una congettura sul parametro e verifico, sulla base del campione, se essa è accettabile (non vera!)

Diversi approcci all'inferenza

- *Classico* (Fisher, Neyman, Pearson): concezione frequentista della probabilità; fa uso unicamente delle informazioni contenute nel campione.
- *Bayesiano* (Lindley, Savage, de Finetti): concezione soggettivista della probabilità; fa uso anche di informazioni a priori sui parametri espresse tramite distribuzioni di probabilità.

- *Teoria delle decisioni* (Wald): tiene conto delle conseguenze di decisioni alternative espresse tramite funzioni di perdita; può fare uso anche di informazioni a priori.

Statistiche campionarie

- *Statistica campionaria*: una qualsiasi funzione

$$g(X_1, X_2, \dots, X_n)$$

del campione. Poichè il campione è casuale la statistica campionaria è una variabile casuale.

- *Statistiche campionarie più comuni*:

- Media campionaria: $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

- Varianza campionaria: $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

- Momento di ordine r : $M_r = \frac{1}{n} \sum_{i=1}^n X_i^r$

- Statistiche d'ordine: $Y_1 \leq Y_2 \leq \dots \leq Y_n$

- Mediana campionaria:

$$Me = \begin{cases} Y_{(n+1)/2} & \text{se } n \text{ è dispari} \\ 0.5(Y_{n/2} + Y_{n/2+1}) & \text{altrimenti} \end{cases}$$

- Campo di variazione: $W = Y_n - Y_1$

Distribuzioni campionarie

- Per una statistica campionaria

$$Y = g(X_1, X_2, \dots, X_n)$$

si ha

$$\begin{aligned} F(y) &= P[g(X_1, X_2, \dots, X_n) \leq y] \\ &= P[(X_1, X_2, \dots, X_n) \in I_y], \end{aligned}$$

dove

$$I_y = \{(x_1, x_2, \dots, x_n) : g(x_1, x_2, \dots, x_n) \leq y\}.$$

Distribuzione della media campionaria

- *In generale*, indipendentemente dalla distribuzione della popolazione, se la media della popolazione è μ e la varianza della popolazione è σ^2 abbiamo che:

$$E(\bar{X}) = \mu, \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

NOTA: la varianza della distribuzione della media campionaria è tanto minore quanto maggior è la dimensione del campione.

Inoltre la varianza è tanto maggiore quanto maggiore è la varianza della popolazione (nel caso limite in cui la popolazione ha varianza nulla, anche la varianza della media campionaria è nulla)

- *Popolazione Bernoulliana:* $X \sim \text{Bin}(1, p)$

$$P(\bar{X} = \bar{x}) = P(n\bar{X} = n\bar{x}) = \binom{n}{n\bar{x}} p^{n\bar{x}} (1-p)^{n-n\bar{x}}$$

$$\bar{X} \sim \frac{1}{n} \text{Bin}(n, p)$$

$$E(\bar{X}) = p, \quad \text{Var}(\bar{X}) = p(1-p)/n$$

- *Popolazione Normale:* $X \sim N(\mu, \sigma^2)$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- *Nel caso di grandi campioni ($n > 30$) :*

$$\lim_{n \rightarrow \infty} P \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z \right) = \Phi(z)$$

dove $\Phi(z)$ è la funzione di ripartizione di una normale standardizzata.

In altre parole la media campionaria, opportunamente standardizzata, si distribuisce asintoticamente come una Normale con media zero e varianza uno.

Questo è il teorema del limite centrale che, enunciato in altri termini dice che la media campionaria si distribuisce, asintoticamente (al crescere dell'ampiezza del campione), come una normale con media uguale alla media della popolazione e varianza pari alla varianza della popolazione divisa per n (ampiezza campionaria):

$$\bar{X} \sim_{n \rightarrow \infty} N(\mu, \sigma^2/n)$$

- **NOTA BENE** ogni combinazione lineare di variabili casuali normali ha ancora una distribuzione normale. Cioè se per esempio $X \sim N(\mu_X, \sigma_X)$ e $Y \sim N(\mu_Y, \sigma_Y)$ ed X è statisticamente indipendente da Y allora

$$aX + bY \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$$

Distribuzione della varianza campionaria

- *In generale*, indipendentemente dalla distribuzione della popolazione, abbiamo che:

$$E(S^2) = \sigma^2$$

$$\text{Var}(S^2) = \frac{\sigma^4}{n} \left(\beta_2 + 2 \frac{n}{n-1} \right)$$

con $\beta_2 = \frac{\mu_4}{\sigma^4} - 3$

- Se poi, in particolare, la popolazione ha

Distribuzione Normale: $X \sim N(\mu, \sigma^2)$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Distribuzione della differenza tra due medie campionarie

- Da due diverse popolazioni provengono i campioni

$$(X_1, X_2, \dots, X_{n_1}) \quad \text{e} \quad (Y_1, Y_2, \dots, Y_{n_2})$$

- $E(\bar{X} - \bar{Y}) = \mu_X - \mu_Y$, $\text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
questo è vero *in generale* indipendentemente dalla
distribuzione della popolazione

- Nel caso particolare di *Popolazioni Normali*:

$$X \sim N(\mu_X, \sigma_1^2), Y \sim N(\mu_Y, \sigma_2^2)$$

$$(\bar{X} - \bar{Y}) \sim N\left(\mu_X - \mu_Y, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

- Nel caso particolare di *Popolazioni Bernoulliane* (n elevato): $X \sim \text{Bin}(1, p_1)$,
 $Y \sim \text{Bin}(1, p_2)$

$$(\bar{X} - \bar{Y}) \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

Distribuzione di rapporti che coinvolgono medie e varianze campionarie

- *Popolazioni Normale:* $X \sim N(\mu, \sigma^2)$

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n - 1)$$

cioè ha una distribuzione T-Student con $(n - 1)$ gradi di libertà

- *Popolazioni Normali:*

$$X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$$

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_c^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t(n_1 + n_2 - 2),$$

con

$$S_c^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

e n_1 ed n_2 le ampiezze dei due campioni rispettivamente