MS 09/10/24

ALCUNE DISTRIBUZIONI MULTIDIMENSIONALI NOTEVOLI.

ESEMPIO

Nati in:

Torino	T	10
Piemonte \ T	P	5
Italia \ P	I	20
Estero	E	2
		37

Questa variabile è qualitativa perché assume valori in un insieme finito non numerico.

Supponiamo ora di selezionare a caso $n = 5$ studenti, per esempio

sample(1:37, 5) in R

Calcolare la probabilità della configurazione

$$(T=2,\ P=0,\ I=3,\ E=0)$$

$$P(T=2,\ P=0,\ I=3,\ E=0) = \frac{\binom{10}{2}\binom{5}{0}\binom{20}{3}\binom{2}{0}}{\binom{37}{5}}$$

Questo è un esempio di calcolo da distribuzione ipergeometrica multidimensionale ($D=2$ fornisce il caso particolare della distribuzione ipergeometrica)

In generale, consideriamo una popolazione di N unità suddivise in D sottopopolazioni di numerosità $N_1 \dots N_d \, N_D$

$d = D-1$ $\qquad N_D = N - \sum_{i=1}^{d} N_i$

e supponiamo di campionare casualmente n unità.

cioè ogni sottoinsieme di cardinalità n è ugualmente probabile.

Siano $X_1 \dots X_d \, X_D$ i conteggi delle rappresentanze delle sottopopolazioni nel campione.
Allora si dice che $(X_1 \dots X_d)$ (o anche $X_1 \dots X_d \, X_D$) ha una distribuzione ipergeometrica (multivariata) con massa di probabilità

$$P(X_1 = x_1, \dots, X_d = x_d, X_D = x_D) = \frac{\binom{N_1}{x_1}\binom{N_2}{x_2}\cdots\binom{N_D}{x_D}}{\binom{N}{n}}$$

dove $\sum_{i=1}^{D} x_i = n$.

Caso particolare: $D=2$

$$P(X = x_1, n - X = n - x_1) =$$
$$= f_{X_1}(x_1) = \frac{\binom{N_1}{x_1}\binom{N-N_1}{n-x_1}}{\binom{N}{n}}$$

e diciamo che X_1 è ipergeometrica (unidimensionale). Anche $n - X_1$ lo è.

$\square$

Nell' esempio, si può parlare di campionamento senza reintroduzione da un'urna (popolazione finita, piccola)

Se invece:

1) il campionamento è fatto con reintroduzione (e.g. generare lettere a caso al computer)

2) la popolazione è molto grande rispetto a n ("infinita") Tecnicamente, $N, N_1, N_2 \ldots N_D \to \infty$

3) c'è un processo generatore di dati che ammette ripetizione casuale

allora la distribuzione di

$$X_1, X_2 \dots X_d, X_D$$

dipende da n e da

p_i = probabilità di selezionare l'i-esimo livello ad ogni prova

Cioè supponiamo che ciascuna prova possa risultare in uno di D livelli e che le prove siano indipendenti. Allora

$$P(X_1=x_1, \dots, X_d=x_d, X_D=x_D) = \overbrace{\frac{n!}{x_1! \dots x_d!\, x_D!}}^{\text{numero di sequenze}} \underbrace{\prod_{i=1}^{D} p_i^{x_i}}_{\text{una particolare sequenza}}$$

dove $\sum_{i=1}^{D} x_i = n$ $\quad \sum_{i=1}^{D} p_i = 1$

Caso particolare: $D=2$

$$P(X_1=x, X_2=n-x) = f_{X_1}(x) = \binom{n}{x} p_1^{x} (1-p_1)^{n-x}$$

Si dice che $(X_1, \dots X_d)$ (oppure a volte $(X_1, \dots X_D)$) ha una distribuzione multinomiale (binomiale se $D=2$). con parametri $n, p_1 \dots p_d$ (e $p_D = 1 - \sum_{i=1}^{d} p_i$)

ESEMPIO

In una analisi sensoriale di un cibo, viene rilevato se il cibo è prevalentemente P (piccante), S (speziato), D (dolce) o A (amaro), con probabilità rispettive p_1, p_2, p_3, p_4. $D=4$.

Se chiedo a 5 valutatori che rispondono in maniera indipendente con quelle probabilità

$$P(\text{"PPADP"}) = p_1 p_1 p_4 p_3 p_1$$

$$= p_1^3 p_2^0 p_3^1 p_4^1$$

$$P(\text{"3P, un A e un D su 5"})$$

$$= \begin{pmatrix} \text{numero} \\ \text{di sequenze} \\ \text{con gli} \\ \text{stessi conteggi} \end{pmatrix} \times p_1^3 p_2^0 p_3^1 p_4^1$$

$$= \frac{5!}{3!\,0!\,1!\,1!} p_1^3 p_2^0 p_3^1 p_4^1$$

$\square$

LA NORMALE MULTIDIMENSIONALE (O MULTIVARIATA)

Definizione in due passi

1) la normale standard
2) la normale generale per trasformazione lineare

1) e' semplice: si definisce normale standard di dimensione k il vettore aleatorio

$$\begin{pmatrix} Z_1 \\ \vdots \\ Z_k \end{pmatrix}$$

con componenti $Z_1, Z_2 \dots Z_k$ i.i.d. $N(0,1)$.

2) dato un vettore $\mu \in \mathbb{R}^n$ e una matrice A di dimensioni $n \times k$ si definisce normale di parametri μ e $\Sigma = AA'$ trasposta il vettore

$$\underset{n\times 1}{X} = \underset{n\times 1}{\mu} + \underset{n\times k\, k\times 1}{AZ}$$

(Gasparini (2013) 2.1.4)

La funzione caratteristica di X

$$\varphi(t) = E(\exp(it'X)) = \exp\left\{ it'\mu - \frac{1}{2} t'\Sigma t \right\}$$

La media (cioè il vettore delle medie delle componenti) di X è

$$E(X) = E(\mu + AZ) = \mu + A\underbrace{E(Z)}_{\text{vettore di } 0} = \mu$$

La matrice di varianze e covarianze di X (cioè la matrice che contiene sulla diagonale $Var(x_i)$ e fuori $Cov(x_i, x_j)$) è

$$VarCov(X) = \begin{bmatrix} Var(x_1) & & & \\ & Var(x_2) & & \\ & Cov(x_i, x_j) & \ddots & \\ & & & Var(x_m) \end{bmatrix}$$

$$= VarCov(\mu + AZ)$$

$$= VarCov(AZ)$$

$$= A \, VarCov(Z) \, A'$$

proprietà di VarCov in generale

$= I$ perché z_i sono i.i.d $N(0,1)$

$$= AA' = \Sigma$$, motivo per cui l'abbiamo chiamata così.

Se A è di rango pieno $m \leq k$ allora con la tecnica dello Jacobiano si dimostra che esiste la densità di X, in particolare

$$f(x) = (2\pi)^{-m/2} (\det \Sigma)^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)' \Sigma^{-1} (x-\mu)\right\}$$

ESEMPIO

$$n = k = 2$$

$$\mu = \begin{pmatrix}0\\1\end{pmatrix}$$

$$A = \begin{pmatrix}1&0\\2&1\end{pmatrix}$$

con due colonne linearmente indipendenti, quindi A è di rango 2 pieno

$\cancel{X}$ $\begin{pmatrix}X\\Y\end{pmatrix}$ notazione solita dall'Analisi 2

$$\begin{pmatrix}X\\Y\end{pmatrix} = \mu + AZ = \begin{pmatrix}0\\1\end{pmatrix} + \begin{pmatrix}1&0\\2&1\end{pmatrix}\begin{pmatrix}Z_1\\Z_2\end{pmatrix}$$

$$= \begin{pmatrix}Z_1\\2Z_1+Z_2+1\end{pmatrix}$$

è una normale con media $\begin{pmatrix}0\\1\end{pmatrix}$ e matrice di varianze e covarianze

$$AA' = \begin{pmatrix}1&2\\2&5\end{pmatrix}$$

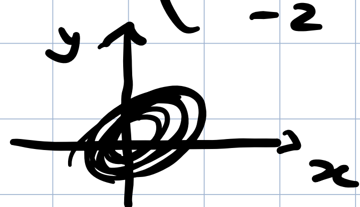
e densità

$$f(x,y) = (2\pi)^{-1}\exp\left\{-\frac{1}{2}\begin{pmatrix}x & y-1\end{pmatrix}\Sigma^{-1}\begin{pmatrix}x\\y-1\end{pmatrix}\right\}$$

dove (esercizio)

$$\Sigma^{-1} = \begin{pmatrix}5&-2\\-2&1\end{pmatrix}$$

e grafo



$\square$