Appunti della Lezione 10

La densita' di probabilita' di Gauss o Normale

Abbiamo gia' visto come trasformare la distribuzione di probabilita' binomiale nel caso in cui sia $p << 1$, $n >> 1$, ovvero la distribuzione di Poisson. Vogliamo ora trovare una soluzione che ci permetta di evitare calcoli complicati (o praticamente impossibili) nel caso in cui la distribuzione binomiale (con $0 < p < 1$) debba essere calcolata per valori grandi di n ($n > 10$). Pensiamo ad esempio al caso in cui vogliamo calcolare la probabilita' di avere piu' di 1100 successi in 2000 lanci di una moneta. Il calcolo binomiale dovrebbe essere

$$B(k > 1100; 2000, 1/2) = \sum_{k=1101}^{2000} B(k; 2000, 1/2)$$

con il primo termine della sommatoria dato da:

$$B(1101; 2000, 1/2) = \binom{2000}{1101} (1/2)^{1101} (1/2)^{2000-1101} = \frac{2000!}{1101!\,899!} (1/2)^{2000}$$

Dovremmo quindi calcolare 900 valori, o essendo furbi "basterebbe" calcolare il valore centrale (k=1000), i 100 valori da k=1001 a 1100, e poi utilizzare l'assioma della certezza per calcolare il valore cercato:

$$B(> 1100; 2000, 1/2) = 1/2(1 - B(1000; 2000, 1/2)) - \sum_{k=1001}^{1100} B(k; 2000, 1/2)$$

Cerchiamo di trovare una soluzione che ci permetta di trasformare la Binomiale in una espressione equivalente, ma piu' facilmente gestibile dal punto di vista analitico.

Utilizziamo un approccio "sperimentale". Immaginiamo di muoverci in una direzione e contemporaneamente di spostarci leggermente a sinistra o a destra di una quantita' ϵ rispetto alla direzione di avanzamento, in modo aleatorio e con una probabilita' di spostarsi a sinistra p ed a destra $1 - p$. Dopo n passi ci troveremo ad una distanza $y = k\epsilon - (n-k)\epsilon = \epsilon(2k-n)$ dalla direzione centrale (k e' il numero di passi a sinistra ed $n - k$ il numero di passi a destra). Lo spostamento laterale segue quindi la statistica Binomiale, e possiamo calcolare lo spostamento medio $\bar{y}$ e la varianza s_y^2 come

$$\bar{y} = E[y] = E[\epsilon(2k - n)] = \epsilon(2\bar{k} - n) = n\epsilon(2p - 1)$$

$$\begin{aligned} s_y^2 = Var[y] &= Var[\epsilon(2k - n)] \\ & Var[2\epsilon k - n\epsilon] \\ & 4\epsilon^2 Var[k] + 0 \\ & (n\epsilon^2) 4p(1 - p) \end{aligned}$$

A questo punto possiamo immaginare di aumentare il numero di passi che facciamo e di rendere lo scartamento quanto piu' piccolo possibile (ovvero

$n \to \infty$ e $\epsilon \to 0$). Nel fare questo vogliamo pero' mantenere intatte le caratteristiche della nostra distribuzione, ovvero evitare che la dispersione della distribuzione si annulli. Questo significa mantenere costante il prodotto $n\epsilon^2$.

Ci apprestiamo a riscrivere la distribuzione discreta $B(k;n,p) \propto f(y)$. Possiamo vedere come l'incremento di un'unita' di k nella Binomiale corrisponda ad un'unita' di 2ϵ nella variabile y ovvero

$$k \to k+1 \Rightarrow y \to y + 2\epsilon$$

e quindi $B(k+1;n,p) \propto f(y+2\epsilon)$. Possiamo ora valutare la seguente espressione

$$\begin{aligned}
\frac{f'(y)}{f(y)} &= \lim_{2\epsilon\to 0} \frac{f(y+2\epsilon)-f(y)}{2\epsilon f(y)} \\
&= \lim_{2\epsilon\to 0} \frac{B(k+1;n,p)-B(k;n,p)}{2\epsilon B(k;n,p)} \\
&= \lim_{2\epsilon\to 0} \frac{1}{2\epsilon} \frac{\binom{n}{k+1}p^{k+1}(1-p)^{n-k-1} - \binom{n}{k}p^k(1-p)^{n-k}}{\binom{n}{k}p^k(1-p)^{n-k}} \\
&= \lim_{2\epsilon\to 0} \frac{1}{2\epsilon} \frac{\frac{p}{k+1} - \frac{1-p}{n-k}}{\frac{1-p}{n-k}} \\
&= \lim_{2\epsilon\to 0} \frac{1}{2\epsilon} \frac{\frac{p(n-k)-(1-p)(k+1)}{(k+1)(n-k)}}{\frac{1-p}{n-k}} \\
&= \lim_{2\epsilon\to 0} \frac{1}{2\epsilon} \frac{pn-pk-k-1+pk+p}{(1-p)(k+1)} \\
&\quad \text{, ma } k = y/2\epsilon + n/2 \\
&= \lim_{2\epsilon\to 0} \frac{1}{2\epsilon} \frac{pn - \frac{y}{2\epsilon} - \frac{n}{2} - 1 + p}{(\frac{y}{2\epsilon}+\frac{n}{2}+1)(1-p)} \\
&= \lim_{2\epsilon\to 0} \frac{1}{2\epsilon} \frac{\frac{2pn\epsilon - y - n\epsilon - 2\epsilon + 2p\epsilon}{2\epsilon}}{\frac{(y+n\epsilon+2\epsilon)(1-p)}{2\epsilon}} \\
&= \lim_{2\epsilon\to 0} \frac{n\epsilon(2p-1) - y - 2\epsilon(1-p)}{2y\epsilon + 2n\epsilon^2 + 4\epsilon^2)(1-p)} \\
&\quad \text{ma } \bar{y} = n\epsilon(2p-1) \quad \text{, e } s_y^2 = n\epsilon^2 4p(1-p) \\
&= \lim_{2\epsilon\to 0} \frac{\bar{y} - y - 2\epsilon(1-p)}{2y\epsilon + 4\epsilon^2)(1-p) + s_y^2/2p} \\
&= \frac{\bar{y}-y}{s_y^2/2p}
\end{aligned}$$

Definiamo ora una nuova quantita' $s^2 = s_y^2/2p$, e procediamo nella risoluzione dell'equazione differenziale (va sottolineato che quando $p = 1/2$ si ha $\bar{y} = 0$ e $s^2 = s_y^2$):

$$\frac{f'(y)}{f(y)} = \frac{\bar{y}-y}{s^2} \to \frac{1}{f(y)}\frac{df(y)}{dy} = -\frac{y-\bar{y}}{s^2} \to \frac{df(y)}{f(y)} = -\frac{y-\bar{y}}{s^2}\,dy$$

ed integrando

$$\int \frac{df(y)}{f(y)} = -\int \frac{y-\bar{y}}{s^2}\,dy \to \ln f(y) = -\frac{(y-\bar{y})^2}{2s^2} + C \to f(y) = Ae^{-(y-\bar{y})^2/2s^2}$$

Per determinare la costante A, basta imporre la normalizzazione della f(y) in modo che essa possa diventare una “densita’ di probabilita‘”:

$$\int_{-\infty}^{\infty} f(y)\,dy = 1 \to A\int_{-\infty}^{\infty} e^{-(y-\bar{y})^2/2s^2}\,dy = 1$$

Dall’analisi sappiamo che esiste il seguente integrale definito

$$\int_0^{\infty} e^{-a^2x^2}\,dx = \frac{1}{2a}\sqrt{\pi}$$

sostituendo $a^2 = 1/2s^2$ e tenendo conto che la funzione e’ simmetrica, si ottiene $A = 1/s\sqrt{2\pi}$ e quindi

$$f(y) = \frac{1}{s\sqrt{2\pi}}\,e^{-(y-\bar{y})^2/2s^2}$$

E’ importante notare che mentre $B(k;n,p)$ e’ una distribuzione di probabilita’ discreta e che quindi i suoi elementi sono probabilita’, la funzione $f(y;\bar{y},s)$ e’ una densita’ di probabilita’ (con dimensione $[y]^{-1}$), che va sotto il nome di densita’ di probabilita’ di Gauss o Normale.

Proprieta’ della densita’ di probabilita’ Normale

Analogamente a quanto visto finora per il caso discreto possiamo determinare i due momenti principali della funzione, ovvero la media e la varianza di una generica densita’ di probabilita’ Normale:

$$E[y] = \mu = \frac{1}{s\sqrt{2\pi}}\int_{-\infty}^{\infty} y\,e^{-(y-\bar{y})^2/2s^2}\,dy$$

sostituendo $x = y - \bar{y}$ otteniamo

$$\begin{aligned}
\mu &= \frac{1}{s\sqrt{2\pi}}\int_{-\infty}^{\infty}(x+\bar{y})\,e^{-x^2/2s^2}\,dx \\
&= \frac{1}{s\sqrt{2\pi}}\int_{-\infty}^{\infty} x\,e^{-x^2/2s^2}\,dx + \frac{1}{s\sqrt{2\pi}}\int_{-\infty}^{\infty}\bar{y}\,e^{-x^2/2s^2}\,dx \\
&\quad \text{il primo integrale e}' \text{ nullo in quanto funzione dispari} \\
&= 0 + \bar{y}\frac{1}{s\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2s^2}\,dx \\
&= 0 + \bar{y}\times 1 \\
\mu &= \bar{y}
\end{aligned}$$

$$Var[y] = \sigma^2 = \frac{1}{s\sqrt{2\pi}}\int_{-\infty}^{\infty}(y-\bar{y})^2\,e^{-(y-\bar{y})^2/2s^2}\,dy$$

sostituendo $x = (y - \bar{y})/s\sqrt{2}$ otteniamo

$$\begin{aligned}
\sigma^2 &= \frac{2s^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} x^2\, e^{-x^2}\, dx \\
&= \frac{2s^2}{\sqrt{\pi}} \frac{\sqrt{\pi}}{2} \\
\sigma^2 &= s^2
\end{aligned}$$

La distribuzione di Gauss e' quindi definita nel seguente modo:

$$N(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{(x-\mu)^2/2\sigma^2}$$

Il problema della distribuzione di Gauss e' la mancanza di una funzione primitiva della gaussiana. Quando vogliamo calcolare la probabilita' discreta, ovvero l'integrale di questa distribuzione tra due valori finiti, dobbiamo ricorrere all'integrazione per via numerica. Per evitare questa operazione, viene definita una particolare densita' di probabilita', che va sotto il nome di *Gaussiana Standard*:

$$G(t;0,1) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-t^2/2}$$

Qualunque distribuzione di Gauss puo' essere ricondotta alla Gaussiana Standard sostituendo a x la variabile standard t con $t = (x-\mu)/\sigma$. La Gaussiana Standard e' integrata per via numerica ed i risultati sono sempre presenti nei libri di statistica, sotto forma di tabelle. In questo modo si puo' calcolare l'integrale di una distribuzione di Gauss, trasformandola in una Gaussiana Standard, ed integrandola tra i corrispondenti valori della distribuzione di Gauss.

$$\begin{aligned}
P_N(x_{min} < x < x_{max}) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{x_{min}}^{x_{max}} e^{(x-\mu)^2/2\sigma^2} \\
P_G(t_{min} < t < t_{max}) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{t_{min}}^{t_{max}} e^{-t^2/2}
\end{aligned}$$

con $t_{min} = (x_{min} - \mu)/\sigma$ e $t_{max} = (x_{max} - \mu)/\sigma$. In pratica la Gaussiana Standard e' una distribuzione in cui la larghezza e' normalizzata in unita' di deviazione standard σ, in particolare

$$P_G(-1 < t < 1) = 0.683 \;;\; P_G(-2 < t < 2) = 0.954 \;;\; P_G(-3 < t < 3) = 0.997$$