MS 16/10/24

SVOLGIMENTO DI ESERCIZI

1. Supponiamo per semplicità che i 40 geni di un genoma possano essere divisi in tre classi disgiunte: classe A di numerosità 20, B e C di numerosità 10 ciascuna. Le tre classi possono per esempio rappresentare funzioni biologiche distinte, o avere origini genetiche diverse.

   (a) si calcoli la probabilità di osservare nessun gene di classe A, due di classe B e tre di classe C in una estrazione casuale di 5 geni;

$$P(N_A=0, N_B=2, N_C=3) = \frac{\binom{20}{0}\binom{10}{2}\binom{10}{3}}{\binom{40}{5}} = \ldots =$$

(dal testo, si evince senza reintroduz. → hyper. mult)

library extraDist

dmvhyper(c(0,2,3),c(20,10,10),5) = 0.008206587

   (b) come approssimazione, si calcoli la stessa probabilità nel caso di campionamento casuale con reintroduzione; (→ multinomiale

mnom

$$\frac{\cancel{5!}^{\,5\cdot 4}}{0!\,2!\,\cancel{3!}} \cancel{\left(\frac{20}{40}\right)^0} \left(\frac{10}{40}\right)^2 \left(\frac{10}{40}\right)^3 =$$

$$= 10/4^5 =$$ dmnom(c(0,2,3),5,c(1/2,1/4,1/4)) = 0.009765625

   (c) si calcolino la probabilità di osservare qualunque campione così estremo o più estremo del precedente rispetto all'ipotesi di campionamento casuale (senza reintroduzione).

   Si noti che la seconda domanda equivale al calcolo del *p-value* per una generalizzazione multivariata di quello che si chiama *test esatto di Fisher* dell'ipotesi nulla di campionamento casuale; a ben guardare, i campioni più estremi sono semplicemente quelli che hanno probabilità più piccole sotto campionamento casuale. Come supporto di calcolo si può usare la libreria `extraDistr` di R, usata anche nella soluzione del primo punto. Si veda infine

   `https://www.youtube.com/watch?v=EF94wPaqXM0`

   per una illustrazione del test di Fisher univariato (una classe per volta) alla *Gene Set Analysis*.

$n=2$

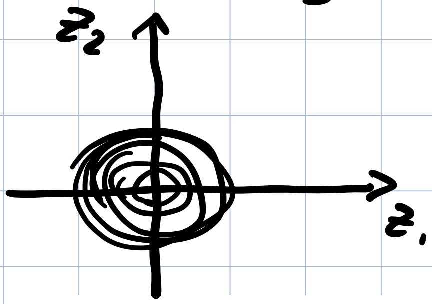
(a) Trovare la distribuzione di: $\begin{pmatrix}X\\Y\end{pmatrix} = \underbrace{\begin{pmatrix}1&0\\0&1/3\end{pmatrix}}_{A}\begin{pmatrix}Z_1\\Z_2\end{pmatrix} \cancel{+\underbrace{\begin{pmatrix}0\\0\end{pmatrix}}_{\mu}}$

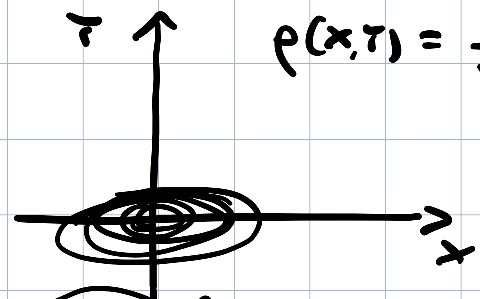
$\Sigma = AA' = \begin{pmatrix}1&0\\0&1/3\end{pmatrix}\begin{pmatrix}1&0\\0&1/3\end{pmatrix} = \begin{pmatrix}1&0\\0&1/9\end{pmatrix}$

$X = Z_1$

$Y = Z_2/3$

$\rho(X,Y) = \frac{0}{\sqrt{1\cdot\frac{1}{9}}} = 0$





(c) Trovare la distribuzione di: $\begin{pmatrix}X\\Y\end{pmatrix} = \underbrace{\begin{pmatrix}1&0\\1&1/3\end{pmatrix}}_{A}\begin{pmatrix}Z_1\\Z_2\end{pmatrix}$

e scrivere $\mathrm{P}(Y < 3X+1)$ come integrale doppio.

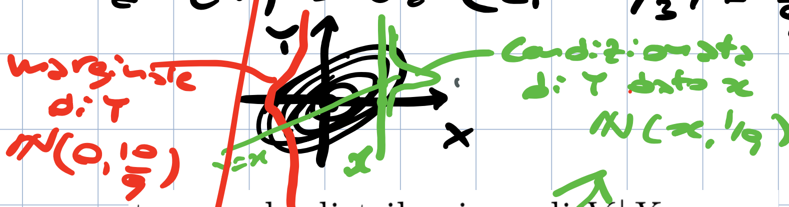
$\Sigma = \text{Var Cov}\begin{pmatrix}X\\Y\end{pmatrix} = \begin{pmatrix}1&0\\1&1/3\end{pmatrix}\begin{pmatrix}1&1\\0&1/3\end{pmatrix} = \begin{pmatrix}1&1\\1&\frac{10}{9}\end{pmatrix}$

$\text{Var}(X) = \text{Var}(Z_1) = 1$

$\text{Var}(Y) = \text{Var}(Z_1 + Z_2/3) = \frac{10}{9}$

$\text{Cov}(X,Y) = 1$

$\rho = \frac{1}{\sqrt{1\times\frac{10}{9}}} = \frac{3}{\sqrt{10}} \approx .95$



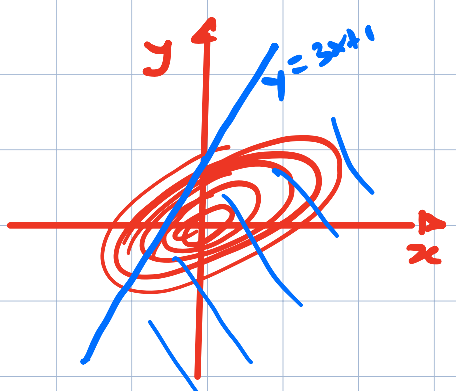
trovare la distribuzione di $Y|X = x$.

$$Y|X=x \sim N\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X), (1-\rho^2)\sigma_Y^2\right)$$

$$= N(x, 1/9)$$

$$P(Y<3X+1) = \int_{-\infty}^{+\infty}\int_{-\infty}^{3x+1} \frac{1}{2\pi}\left(\left(1-\frac{9}{10}\right)\frac{10}{9}\right)^{-\frac{1}{2}} \times$$

$$\times \exp\left\{-\frac{1}{2(1-\frac{9}{10})}\left[x^2 - 2\frac{3}{\sqrt{10}}\,x\,\frac{y}{\sqrt{\frac{10}{9}}} + \frac{y^2}{\frac{10}{9}}\right]\right\} dy\, dx$$



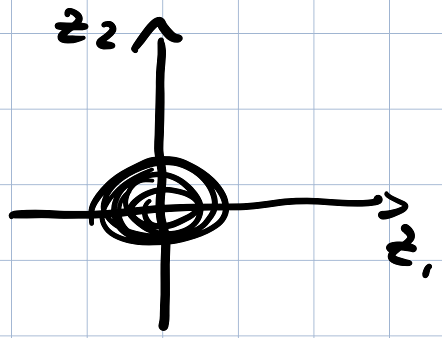
$= \dots$ un po' di Analisi Numerica oppure Montecarlo

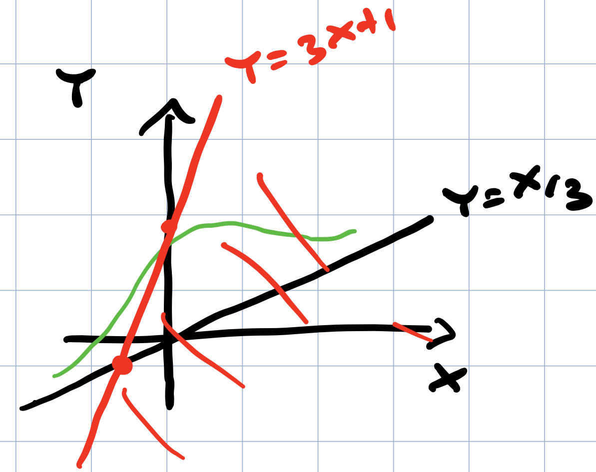
(b) Trovare la distribuzione di: $\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1/3 & 0 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$

e calcolare $\mathrm{P}(Y < 3X + 1)$.

$X = Z_1$

$Y = Z_1/3 = X/3$





$\begin{pmatrix} X \\ Y \end{pmatrix}$ non ha densità rispetto alla misura di Lebesgue in $\mathbb{R}^2$ perché vive su una retta (che ha dimensione 1)

$$P(Y < 3X + 1)$$

$$= P\left(X > -\tfrac{3}{8}\right)$$

$$= 1 - P\left(Z < -\tfrac{3}{8}\right)$$

$$= P\left(Z > \tfrac{3}{8}\right) \approx 0.6462$$

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1/3 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1/3 \\ 0 & 0 \end{pmatrix} \right)$$

3. Siano $Z_1, Z_2, \dots Z_3$ i.i.d. normali standard e

$$\begin{pmatrix} X \\ Y \\ S \end{pmatrix} = \overset{A}{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 1 & 1 \end{pmatrix}} \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} + \overset{\mu}{\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}}$$

Σ_{11}

(a) Calcolare media e matrice di varianze e covarianze di $(X, Y, S)'$.

(b) Scrivere $\mathrm{P}(X < 0 \cap S < 0)$ come integrale doppio.

(c) Trovare la distribuzione condizionata di S dato $X = x \cap Y = y$.

(a)

$$E\begin{pmatrix} X \\ Y \\ S \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

$$\mathrm{VarCov}\begin{pmatrix} X \\ Y \\ S \end{pmatrix} = AA' = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix}$$

(b)

$$\begin{pmatrix} X \\ S \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix}\right)$$

stessa cosa di 2(c)

$$P(X<0 \cap S<0) = \int_{-\infty}^{0}\int_{-\infty}^{0} f_{(X,S)}(x,s)\, dx\, ds$$

(c) applico la formula più generale che coinvolgerà

$$\Sigma_{11}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1/4 \end{pmatrix}$$

otteniamo

$$S \mid X=x, Y=y \sim \mathcal{N}\left(x + \frac{y}{2} - \frac{1}{2}, 1\right)$$