Regressione

Il problema della regressione è cercare di capire che relazione c'è tra alcuni dati che controllo in entrata e l'uscita che misuro con un certo errore.

REGRESSIONE

(VARIABILI IN INGRESSO)
DATI IN ENTRATA $X_1, X_2 \dots X_N$

VALORE IN USCITA $Y = f(X_1, \dots X_N)$
(VARIABILE IN USCITA)

Il problema è determinare f. Può essere necessario determinarne la forma (polinomio, seno, coseno, logaritmo…) oppure possiamo immaginare che ci sia una teoria più o meno consolidata in cui si dice che Y dipende in una certa maniera dalle variabili in ingresso, si conosce la forma di f ma non tutti i coefficienti che compongono f.

Problema: determinare f come forma
oppure i coeff. di f

REGRESSIONE LINEARE: f funzione lineare
di $X_1, \dots, X_N$

REGRESSIONE LINEARE SEMPLICE: $N=1$ ovvero
una sola variabile
in ingresso.

In questa categoria rientrano tutti gli studi di dipendenza dalle variabili ambientali rispetto a una certa caratteristica degli individui di una popolazione. Per esempio, dal punto di vista della fisica la dipendenza della velocità di un grave in caduta dall'accelerazione di gravità.

L'idea è di fare un esperimento ripetuto più volte in corrispondenza di valori diversi della variabile d'ingresso, misurare i valori in uscita e cercare di trovare una retta nel caso lineare o una curva nel caso non lineare, in grado di interpolare al meglio i dati osservati.

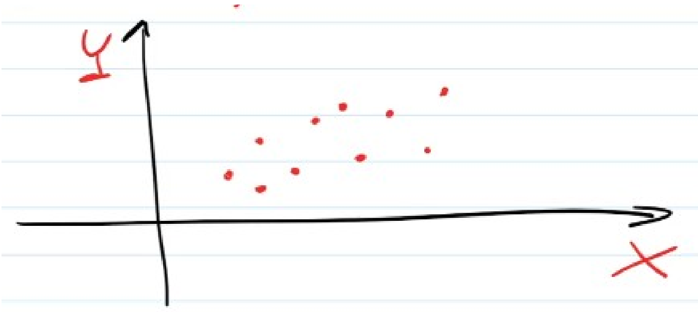
Regressione lineare semplice

REGRESSIONE LINEARE SEMPLICE

IN TEORIA $Y = \alpha + \beta X$ con α e β parametri incogniti

Nel mondo reale, Y è affetto da un errore, proprio a causa di questo errore di misurazione non è facile determinare alfa e beta.

IN PRATICA $Y = \alpha + \beta X + \text{errore}$



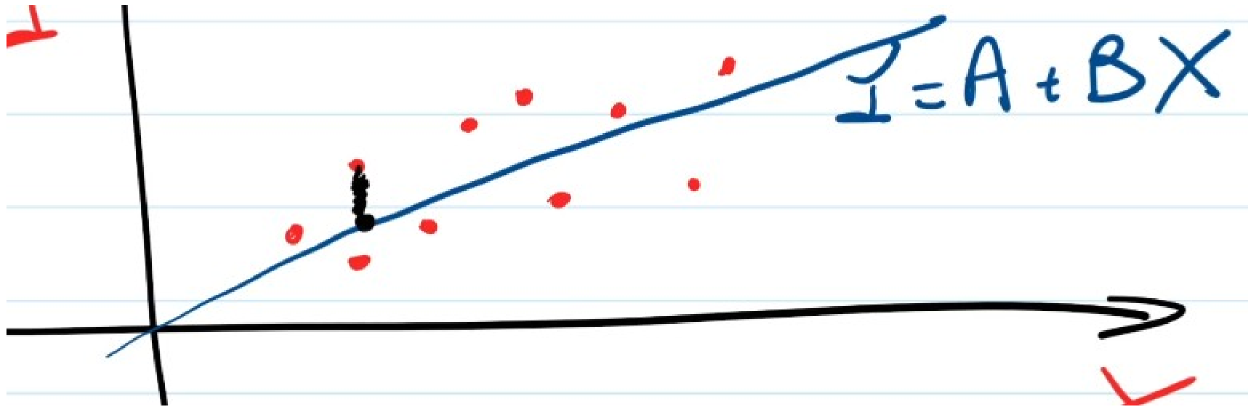
valori di Y per X_K fissato $(X = X_K)$

Y_K	X_K
Y_1	X_1
Y_2	X_2
	$\vdots$
Y_M	X_M

Come scegliere la migliore retta, quella che interpola meglio i dati? Metodo dei minimi quadrati o regressione. Usiamo due stimatori per stimare alfa e beta.

STIMATORI A per stimare α
B per stimare β

Vuol dire passare da parametri numerici a variabili casuali. Come sceglierli? Dall'idea di calcolarsi la differenza che c'è tra la retta Y = A + BX e i punti dei dati.



Il puntino nero è il valore teorico dello stimatore, quello rosso è il valore misurato, la distanza si considera al quadrato e viene sommata alle altre distanze.

Queste distanze o si considerano come differenze (positive o negative) oppure usando il valore assoluto, oppure fare la somma dei quadrati (la cosa più semplice dal punto di vista matematico).

$$SS^2 = \sum_{k=1}^{M} \left(Y_k - (A + BX_k)\right)^2$$

somma dei quadrati delle distanze tra valore misurato e punto $\in$ retta

SS quadro è la somma dei qudrati (in inglese sum e square). Scegliamo A e B affinche la somma dei quadrati sia la più piccola possibile. Minimizziamo la somma dei quadrati rispetto ad A e B.

A e B sono scelti in modo da minimizzare SS^2

Per minimizzare una funzione che dipende da due variabili, rispetto ad A e B, bisogna derivare rispetto ad A e porla a 0, poi la derivata della stessa funzione rispetto a B e porla a 0.

$$\begin{cases} \dfrac{\partial SS^2}{\partial A} = 0 \\ \dfrac{\partial SS^2}{\partial B} = 0 \end{cases} \Rightarrow \begin{cases} \sum_{k=1}^{M} 2\left(Y_k - A - BX_k\right)(-1) = 0 \\ \sum_{k=1}^{M} 2\left(Y_k - A - BX_k\right)(-X_k) = 0 \end{cases}$$

$$\begin{cases} \sum_{k=1}^{M} Y_k - \overset{\shortparallel}{\sum_{k=1}^{M} A} - B \sum_{k=1}^{M} X_k = 0 \\ \sum_{k=1}^{M} Y_k X_k - A \sum_{k=1}^{M} X_k - B \sum_{k=1}^{M} X_k^2 = 0 \end{cases}$$

NOTAZIONE $\overline{X} = \sum_{k=1}^{M} \frac{X_k}{M}$; $\overline{Y} = \sum_{k=1}^{M} \frac{Y_k}{M}$

Non si parla di media campionaria, perché per motivi diversi la media aritmetica delle Xk e delle Yk non sono due grandezze che rappresentano delle medie campionarie.

$$\begin{cases} \sum_{k=1}^{M} Y_k - \underbrace{\sum_{k=1}^{M} A}_{=AM} - B \sum_{k=1}^{M} X_k = 0 \\ \sum_{k=1}^{M} Y_k X_k - A \sum_{k=1}^{M} X_k - B \sum_{k=1}^{M} X_k^2 = 0 \end{cases}$$

DIVIDO (1) per M $\Rightarrow$ $\overline{Y} - A - B\overline{X} = 0$

$A = \overline{Y} - B\overline{X}$

$$\sum_{k=1}^{M} Y_k X_k - A \underbrace{\sum_{k=1}^{M} X_k}_{=\overline{X}M} - B \sum_{k=1}^{M} X_k^2 = 0 \quad (2)$$

$$(2) \Rightarrow \sum_{k=1}^{M} Y_k X_k - M\overline{X}(\overline{Y} - B\overline{X}) - B \sum_{k=1}^{M} X_k^2 = 0$$

$$\sum_{k=1}^{M} Y_k X_k - M\bar{Y}\bar{X} = B\left(\sum_{k=1}^{M} X_k^2 - M\bar{X}^2\right)$$

$$B = \frac{\sum_{k=1}^{M} Y_k X_k - M\bar{Y}\bar{X}}{\sum_{k=1}^{M} X_k^2 - M\bar{X}^2}$$

Quindi:

$$\begin{cases} B = \dfrac{\sum_{k=1}^{M} Y_k X_k - M\bar{Y}\bar{X}}{\sum_{k=1}^{M} X_k^2 - M\bar{X}^2} \\ A = \bar{Y} - B\bar{X} \end{cases}$$

Mettiamoci in condizioni particolari.

Immagino di avere un controllo di precisione estremo sui valori della variabile di ingresso. Posso far variare la variabile d'ingresso come mi pare con un errore nullo, è una variabile deterministica, la vario a mio piacere per fare esperimenti senza errore. La variabile d'uscita invece da errore, è una variabile casuale, suppongo sia gaussiana, che la Y abbia sempre varianza sigma quadro, ma la media sia proprio quella teorica.

IPOTESI

1) X é una quantità che posso controllare praticamente senza errore $\Rightarrow$ le $X_1, \dots X_M$ sono dei valori scelti da chi fa l'esperimento e NON sono v.c.

Come facciamo a capire che A e B sono minimi e non massimi? Posso considerare una retta distante anni luce dalla nuvola di punti, ma ce ne può essere un'altra ancora più lontana. Quindi la somma di quadrati può presentare un minimo ma non un massimo.

$$2)\ Y_K \sim N(\alpha + \beta X_K, \sigma^2)$$

l'idea alla base di questa ipotesi è che

$$Y_K = \alpha + \beta X_K + \text{errore} \quad \text{con errore} \sim N(0, \sigma^2)$$

Se sommiamo a una gaussiana del genere dei numeri, si ottiene una gaussiana con media traslata.

$$Y_K = \overbrace{\gamma + \text{errore}} \sim N(\gamma, \sigma^2)$$

Se non ci fosse nessun errore nella variabile d'uscita varrebbe la formula sopra senza errore, ma ci sono vari fattori che lo impediscono.

$$3)\ Y_1, \dots Y_M \text{ SONO V.C. INDIPENDENTI}$$

In alcuni casi possiamo dire che sia naturale, in altri casi no.

Queste ipotesi ci dicono molto di più sugli stimatori introdotti prima.

$$B = \frac{\sum_{K=1}^{M} Y_K X_K - M\bar{Y}\bar{X}}{\sum_{K=1}^{M} X_K^2 - M\bar{X}^2}$$

B è una variabile casuale perché le Yk sono variabili casuali.

$$M\bar{Y} = \sum_{K=1}^{M} Y_K$$

$$M\bar{Y}\bar{X} =$$

$$= \frac{\sum_{K=1}^{M} \left[Y_K (X_K - \bar{X})\right]}{\sum_{K=1}^{M} X_K^2 - M\bar{X}^2}$$

$$= \frac{\sum_{k=1}^{M} \left[Y_k (x_k - \bar{x}) \right]}{\sum_{k=1}^{M} x_k^2 - M\bar{x}^2}$$

La parte evidenziata è un coefficiente numerico.

B è una combinazione
lineare delle Y_k

Siccome le Yk sono gaussiane e indipendenti, B si comporta come una gaussiana.

Dato che le Y_k sono v.c. gaussiane e sono
indipendenti per la riproducibilità
$B \sim N$

Che media ha B? Per essere uno stimatore corretto, deve avere media beta.

$$E[B] = E\left[\frac{\sum_{k=1}^{M} \left[(x_k - \bar{x}) Y_k \right]}{\sum_{k=1}^{M} x_k^2 - M\bar{x}^2} \right]$$

$$= \frac{\sum_{k=1}^{M} \left[(x_k - \bar{x}) E[Y_k] \right]}{\sum_{k=1}^{M} x_k^2 - M\bar{x}^2}$$

$$= \frac{\sum_{k=1}^{M} \left[(x_k - x)(\alpha + \beta x_k) \right]}{\sum_{k=1}^{M} x_k^2 - M\bar{x}^2}$$

$$= \alpha \frac{\sum_{k=1}^{M}(x_k - \bar{x})}{\sum_{k=1}^{M} x_k^2 - M\bar{x}^2} + \beta \frac{\sum_{k=1}^{M}(x_k^2 - \bar{x}x_k)}{\sum_{k=1}^{M} x_k^2 - M\bar{x}^2}$$

$$= \alpha \frac{\left[\overbrace{\left(\sum_{k=1}^{M} x_k\right)}^{M\bar{x}} - M\bar{x}\right]}{\sum_{k=1}^{M} x_k^2 - M\bar{x}^2} + \beta \frac{\left(\left(\sum_{k=1}^{M} x_k^2\right) - \bar{x}\overbrace{\sum_{k=1}^{M} x_k}^{M\bar{x}}\right)}{\sum_{k=1}^{M} x_k^2 - M\bar{x}^2}$$

$$= \cancel{\alpha \frac{\left[\underbrace{\overbrace{\left(\sum_{k=1}^{M} x_k\right)}^{M\bar{x}} - M\bar{x}}_{=0}\right]}{\sum_{k=1}^{M} x_k^2 - M\bar{x}^2}} + \beta \frac{\left(\left(\sum_{k=1}^{M} x_k^2\right) - \bar{x}\overbrace{\sum_{k=1}^{M} x_k}^{M\bar{x}}\right)}{\sum_{k=1}^{M} x_k^2 - M\bar{x}^2} =$$

$$= \beta \frac{\left(\sum_{k=1}^{M} x_k^2 - M\bar{x}^2\right)}{\sum_{k=1}^{M} x_k^2 - M\bar{x}^2} = \beta \Rightarrow E[B] = \beta$$

B è corretto!

Sappiamo che B è una gaussiana che dipende da Yk e ha media beta.

$$A = \bar{Y} - B\bar{x}$$

$$E[A] = E[\bar{Y} - B\bar{x}] = E[\bar{Y}] - E[B]\bar{x} =$$

$$= E\left[\sum_{k=1}^{M} \frac{Y_k}{M}\right] - \beta\bar{x} = \frac{1}{M}\sum_{k=1}^{M} E[Y_k] - \beta\bar{x} =$$

$$= \frac{1}{M}\sum_{k=1}^{M}(\alpha+\beta X_k) - \beta\overline{X} =$$

$$= \frac{1}{M}\left\{\sum_{k=1}^{M}\alpha + \beta\sum_{k=1}^{M}X_k\right\} - \beta\overline{X} = \frac{\cancel{M}\alpha}{\cancel{M}} + \cancel{\beta\frac{\sum_{k=1}^{M}X_k}{M}} \cancel{- \beta\overline{X}}$$

$$\frac{\sum_{k=1}^{M}X_k}{M} = \overline{X}$$

$$= \alpha \quad \Rightarrow E[A] = \alpha \quad A \text{ è stimatore corretto per } \alpha$$

Facciamo la varianza di B.

$$Var(B) = Var\left(\frac{\sum_{k=1}^{M}(X_k-\overline{X})Y_k}{\sum_{k=1}^{M}X_k^2 - M\overline{X}^2}\right) \underset{\substack{\text{INDIPENDENZA} \\ \text{DELLE } Y_K}}{=}$$

$$= \frac{\sum_{k=1}^{M}(X_k-\overline{X})^2\, Var(Y_k)}{\left(\sum_{k=1}^{M}X_k^2 - M\overline{X}^2\right)^2}$$

$$= \sigma^2 \frac{\sum_{k=1}^{M}(X_k-\overline{X})^2}{\left(\sum_{k=1}^{M}X_k^2 - M\overline{X}^2\right)^2}$$

Facciamo una piccola parentesi.

$$\circledast \quad \sum_{k=1}^{M}(X_k-\bar{X})^2 = \sum_{k=1}^{M}\left(X_k^2+\bar{X}^2-2X_k\bar{X}\right) =$$

$$= \sum_{k=1}^{M}X_k^2 + M\bar{X}^2 - 2\bar{X}\underbrace{\sum_{k=1}^{M}X_k}_{=\bar{X}M} =$$

$$= \sum_{k=1}^{M}X_k^2 + M\bar{X}^2 - 2M\bar{X}^2 = \sum_{k=1}^{M}X_k^2 - M\bar{X}^2$$

In altre parole, sostituisco il denominatore con la sommatoria evidenziata.

$$\sigma^2 \, \frac{\sum_{k=1}^{M}(X_k-\bar{X})^2 \; \circledast}{\left(\sum_{k=1}^{M}X_k^2 - M\bar{X}^2\right)^2} = \sigma^2 \, \frac{\sum_{k=1}^{M}(X_k-\bar{X})^2}{\left(\sum_{k=1}^{M}(X_k-\bar{X})^2\right)^2}$$

$$= \frac{\sigma^2}{\sum_{k=1}^{M}(X_k-\bar{X})^2} \Rightarrow Var(B) = \frac{\sigma^2}{\sum_{k=1}^{M}(X_k-\bar{X})^2}$$

Il significato è poco immediato, ma noi vogliamo una varianza molto piccola per B, affinché sia un buon stimatore. Per avere una varianza piccola, il denominatore deve essere grande.

N.B. nei passaggi precedenti:

$$Var(B) = Var\left(\frac{\sum_{k=1}^{n}(X_k-\bar{X})Y_k}{\sum_{k=1}^{M}X_k^2 - M\bar{X}^2}\right) = Var\left(\sum_{k=1}^{M}\gamma_k Y_k\right) \underset{INDIP.}{=} \sum_{k=1}^{n} Var(\gamma_k Y_k) =$$

$$= \sum_{k=1}^{n}\gamma_k^2 \, Var(Y_k) = *$$

$$* = \frac{\sum_{k=1}^{M}(X_k-\bar{X})^2 \, Var(Y_k)}{\left(\sum_{k=1}^{M}X_k^2 - M\bar{X}^2\right)^2} =$$

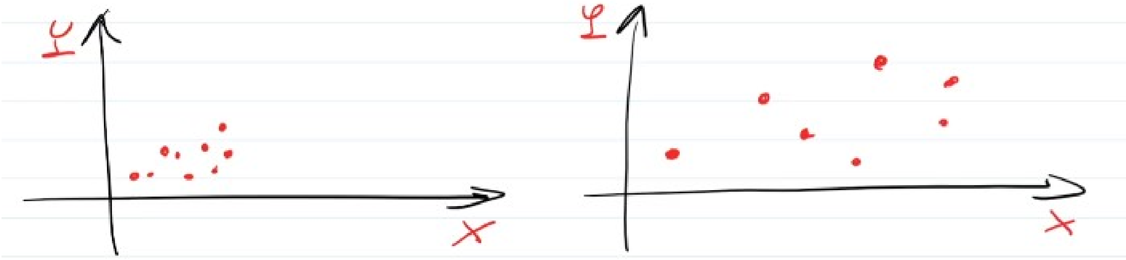
$$\hookrightarrow Var(Y_k) = \sigma^2$$

E anche:

$$= \sigma^2 \frac{\sum_{k=1}^{M}(x_k-\bar{x})^2}{\left(\sum_{k=1}^{M}x_k^2 - M\bar{x}^2\right)^2} \circledast = \sigma^2 \frac{\cancel{\sum_{k=1}^{M}(x_k-\bar{x})^2}}{\left(\sum_{k=1}^{M}(x_k-\bar{x})^2\right)^{\cancel{2}}} = \frac{\sigma^2}{\sum_{k=1}^{M}(x_k-\bar{x})^2}$$

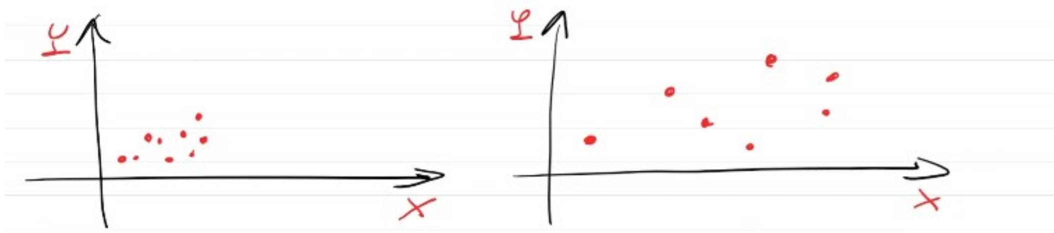
Per ottenere $Var(B)$ piccola a σ^2 fissato, occorre avere $\sum_{k=1}^{M}(x_k-\bar{x})^2 \gg 1$

Immaginiamo queste due situazioni.



Nel primo caso, i valori di X sono molto vicini tra di loro, quindi molto vicini alla loro media aritmetica, dunque Xk – la media aritmetica tutto al quadrato sarà una quantità molto piccola.

Nel secondo caso, i valori di Xk sono stati presi in un range molto ampio, ci sarà qualche X vicino alla media, ma altri no. Siccome si fa Xk – la media aritmetica al quadrato, avremo questa quantità grande.



In questo caso x_k tutti vicini $\Rightarrow \sum_{k=1}^{M}(x_k-\bar{x})^2$ "piccolo"

In questo caso x_k tutti distanti $\Rightarrow \sum_{k=1}^{M}(x_k-\bar{x})^2$ "grande"

Dal punto di vista grafico, se si vuole scegliere la retta che interpola meglio la rete di punti è più semplice determinare con una buona precisione il coefficiente di una retta su dei dati sparpagliati o più vicini? Quelli più vicini sono più complicati perché essendoci dati molto vicini è più difficile stimare il coefficiente che interpola i dati. Se invece i dati sono molto distanti, delle piccole differenze del coefficiente angolare si osservano subito come effetto sull'interpolazione dei dati. Questo si racchiude nella varianza: tanto è più piccola quanto più inizialmente scegliamo dei valori di Xk distanti tra gli altri; tanto è più grande se lavoro con valori di Xk molto vicini tra di loro. È un'interpretazione geometrica dell'effetto. La scelta degli Xk incide nella varianza dello stimatore B.

In questo caso
X_k tutti vicini
$\Rightarrow \sum_{k=1}^{M}(x_k-\bar{x})^2$ "piccolo"
più complesso stimare il coeff. angolare della retta che interpola i dati

In questo caso
X_k tutti distanti
$\Rightarrow \sum_{k=1}^{M}(X_k-\bar{X})^2$ "grande"
è più semplice stimare il coeff. angolare della retta che interpola i dati

Se consideriamo le Xk (non variabili casuali) e li prendiamo molto diversi gli dagli altri, otteniamo una nuvola di punti più facile da interpolare dal punto di vista dei coefficienti angolari. Al denominatore della varianza di B c'è:

$$\sum_{k=1}^{M}(X_k-\bar{X})^2$$

È al denominatore, quindi la varianza diventa molto grande se i valori sono vicini e viceversa.

OSS. Modellare le X_k come quantità deterministiche è ovviamente un'approssimazione molto "forte" dettata dall'esigenza di semplificare i calcoli

Non è ovvio misurare le Xk se fossero state variabili casuali, diventa un calcolo molto complicato. I numeri precisissimi si hanno per l'esigenza di semplificare l formulе. L'errore c'è ma è trascurabile.

È ragionevole supporre le X_K con errore trascurabile rispetto a quello delle Y_K per via delle semplificazioni del modello $Y = \alpha + \beta X$ (Y potrebbe dipendere da variabili non prese in esame il cui effetto è descritto da un errore casuale)

Naturalmente, possiamo passare da una regressione lineare semplice a una regressione lineare multipla, cioè si considerano più variabili da cui Y dipende in maniera lineare. Andrò a scrivere delle specie di matrici che permettono di determinare l'espressione esplicita degli stimatori dei coefficienti lineari.

N.B. $\overline{X} = \sum_{K=1}^{M} \frac{X_K}{M}$ NON è una media campionaria perché le X_K sono dei valori, dei numeri NON delle v.c.

$\overline{Y} = \sum_{K=1}^{M} \frac{Y_K}{M}$ NON è una media campionaria

Non è una media campionaria perché le Yk non provengono dalla stessa popolazione, sono tutte gaussiane ma con medie diverse.

$\overline{Y} = \sum_{K=1}^{M} \frac{Y_K}{M}$ NON è una media campionaria perché le Y_K NON provengono dalla stessa popolazione dato che presentano il parametro $\mu_K = \alpha + \beta X_K$ della distribuzione che varia al variare di K

Restano notazioni di medie aritmetiche, ma non campionarie.

Esercizio

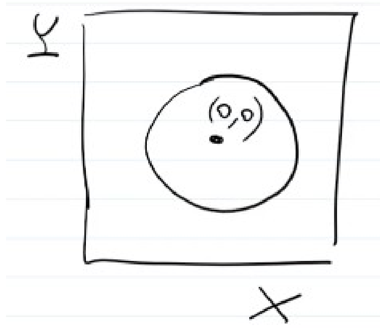
es.

$$X, Y \sim N(0, \sigma^2) \quad \text{INDIPENDENTI}$$

$$R = \sqrt{X^2 + Y^2}$$

X e Y potrebbero essere le coordinate di arrivo di una freccetta sul bersaglio. Si immaginano X e Y gaussiane (tante imprecisioni) di media nulla e varianza sigma quadro. Si spera arrivi a 0, ma c’è un errore dovuto da sigma quadro. Quindi è la distanza dal centro del bersaglio.

Si chiede di determinare la funzione di ripartizione di R.



$$F_R(a) \; ?$$

È un esercizio complicato.

$$F_R(a) = P(R \leq a)$$

$$F_R(a) = P(R \leq a) = \begin{cases} 0 & \text{se } a < 0 \\ ? & \text{se } a \geq 0 \end{cases}$$

se $a \geq 0$

$$F_R(a) = P(R \leq a) = P\left(\sqrt{X^2 + Y^2} \leq a\right)$$

$$) = \iint_{\sqrt{x^2+y^2} \le a} f(x,y)\,dx\,dy$$

$$= \iint_{\sqrt{x^2+y^2} \le a} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{y^2}{2\sigma^2}}\,dx\,dy$$

$\sqrt{x^2+y^2} \le a$

$$= \frac{1}{\sigma^2 2\pi} \iint_{\sqrt{x^2+y^2} \le a} e^{-\frac{(x^2+y^2)}{2\sigma^2}}\,dx\,dy$$

$$= \left|\begin{array}{l} \text{COORDINATE POLARI} \\ x = \rho\cos\theta \\ y = \rho\sin\theta \\ x^2+y^2 = \rho^2 \\ dx\,dy = \rho\,d\rho\,d\theta \\ \theta \in [0, 2\pi] \\ \rho \in [0, a] \end{array}\right.$$

$$= \frac{1}{\sigma^2 2\pi} \int_0^{2\pi} \left(\int_0^a e^{-\frac{\rho^2}{2\sigma^2}} \rho\,d\rho \right) d\theta$$

$$= \frac{1}{\sigma^2 2\pi}\, 2\pi \left[-e^{-\frac{\rho^2}{2\sigma^2}} \sigma^2 \right]_0^a$$

$$= \frac{1}{\cancel{\sigma^2 2\pi}} \cancel{2\pi} \left[-e^{-\frac{\rho^2}{2\sigma^2}} \cancel{\sigma^2} \right]_0^a = 1 - e^{-\frac{a^2}{2\sigma^2}}$$

$$F_R(a) = \begin{cases} 0 & \text{se } a < 0 \\ 1 - e^{-\frac{a^2}{2\sigma^2}} & \text{se } a \geq 0 \end{cases}$$

Cos'è importante di questo esercizio? Il fatto che comunque tutte le volte che si ha una coppia di variabili casuali continue la probabilità che stiano in una certa regione del piano cartesiano, si calcola facendo un integrale doppio.

DIMOSTRARE CHE IL VALORE PIÙ PROBABILE DI R NON È 0

Per valore più probabile di una v.c. continua intendo il massimo della funzione di densità di prob.

$$f_R(a) = \frac{d F_R(a)}{da} = \begin{cases} 0 & \text{se } a < 0 \\ -e^{-\frac{a^2}{2\sigma^2}} \left(-\frac{\cancel{2}a}{\cancel{2}\sigma^2} \right) = \frac{a}{\sigma^2} e^{-\frac{a^2}{2\sigma^2}} & \text{se } a \geq 0 \end{cases}$$

RICERCA DEI MASSIMI RELATIVO

$$\frac{d f_R(a)}{da} = 0 \Rightarrow \frac{1}{\sigma^2} e^{-\frac{a^2}{2\sigma^2}} + \frac{a}{\sigma^2}\left(+ e^{-\frac{a^2}{2\sigma^2}} \left(-\frac{2a}{2\sigma^2} \right) \right) = 0$$

$$\frac{e^{-\frac{a^2}{2\sigma^2}}}{\sigma^2} - \frac{a^2}{\sigma^4} e^{-\frac{a^2}{2\sigma^2}} = 0$$

$$\frac{e^{-\frac{a^2}{2\sigma^2}}}{\sigma^4} \left(\sigma^2 - a^2 \right) = 0 \Rightarrow a^2 = \sigma^2$$

$$a = \sigma$$

Il valore più probabile è in a = sigma. Non ci troviamo più nel caso discreto, ma consideriamo il massimo della funzione di densità.

Metodo per la simulazione di variabili casuali a partire da U variabile casuale uniforme (0, 1)

METODO PER LA SIMULAZIONE DI V.C. A PARTIRE DA $U \sim U(0,1)$

X v.c. continua con $F_X(a)$

Escludiamo gli intervalli in cui la funzione di densità è nulla. Concentriamoci dove la funzione di ripartizione non è costante,

$$Y = F_X^{-1}(U) \quad \text{v.c.}$$

$$F_Y(a) = P(Y \leq a) = P(F_X^{-1}(U) \leq a)$$

La Y è data dalla funzione inversa della distribuzione della X con argomento U.

$$= P(U \leq F_X(a))$$

$$\int_0^{F_X(a)} 1 \, du = F_X(a)$$

Y è proprio la variabile casuale X che volevamo simulare.

$$\Rightarrow X = Y \Rightarrow X = F_X^{-1}(U)$$

Il metodo della funzione inversa (questo illustrato) viene usato tutte le volte che conosco la funzione di ripartizione esplicita e la posso invertire. Non funziona con la gaussiana perché non abbiamo la primitiva. Con tutte le altre variabili con funzione di ripartizione conosciuta si può usare questa tecnica per simulare la X con variabili random.

es. $X \sim E(\lambda)$

$$F_X(a) = \begin{cases} 0 & \text{se } a < 0 \\ 1 - e^{-\lambda a} & \text{se } a \geq 0 \end{cases}$$

$F_X^{-1}(y)$: $\quad y = 1 - e^{-\lambda a} \Rightarrow 1-y = e^{-\lambda a}$

$$\ln(1-y) = -\lambda a$$

$$a = -\frac{1}{\lambda}\ln(1-y)$$

$$F_X^{-1}(y) = -\frac{1}{\lambda}\ln(1-y)$$

$$X = -\frac{1}{\lambda}\ln(1-U)$$

N.B. $1-U \sim U(0,1)$

$$\Downarrow$$

$$X = -\frac{1}{\lambda}\ln(U)$$