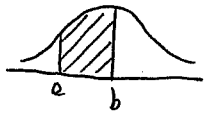
Richiami di statistica I sulle gaussiane standard

- Gaussiana canonica (normale standard) $N(0,1)$

densità $\longrightarrow$ $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ [pdF]

$$P(X \in [a,b]) = \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$



con X v.a. che ha la densità $f(x)$

- Gaussiane qualsiasi $N(\mu, \sigma^2)$

densità $\longrightarrow$ $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ [pdF] con $\mu = E[X]$ = media, $\sigma^2 = Var[X]$ = varianza

- Legame tra $N(\mu, \sigma^2)$ e $N(0,1)$

Data $X \sim N(\mu, \sigma^2)$, allora (per un teorema) $X = \sigma Z + \mu$, con $Z \sim N(0,1)$

N.B.: la moltiplicazione per σ è una dilatazione, mentre l'aggiunta di μ è una traslazione

dim: $\frac{X-\mu}{\sigma} = Z$ (standardizzazione) $\longrightarrow$ Z è gaussiana con $E[Z] = \frac{E[X]-\mu}{\sigma} = 0$ (è un'operazione lineare)

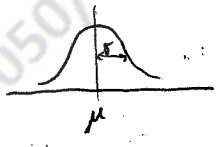
$Var[Z] = \frac{Var[X-\mu]}{\sigma^2} = \frac{Var[X]}{\sigma^2} = 1$ (la varianza è invariante per traslazioni)

$\Big\} N(0,1)$

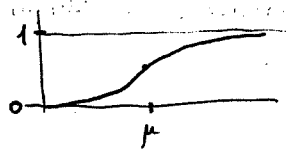
- Funzione cumulativa (o funzione di ripartizione)

$$F(t) = P(X \le t) = \int_{-\infty}^{t} f(x)\, dx \quad [cdF]$$

la pdf di una $X \sim N(\mu, \sigma^2)$ è



la cdf di una $X \sim N(\mu, \sigma^2)$ è

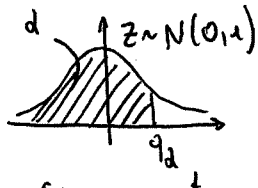


MASTER COPY
Tel. 388/9837745

- quantile

① Definizione di quantile gaussiano standard

Si chiama quantile gaussiano di ordine α il numero q_α tale che $\Phi(q_\alpha) = \alpha$

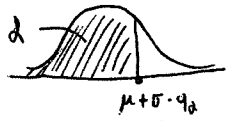


con α assegnata

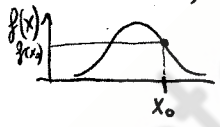
Quantile e cdf sono operazioni inverse: nella cdf fisso un punto t e calcolo l'area, mentre nel quantile fisso l'area e trovo un punto.

② Comandi di R

- `pnorm(t, μ, σ)` Calcola la cdf (funzione cumulativa) di una $X \sim N(\mu, \sigma^2)$, nell'intervallo tra $-\infty$ e t. Cioè calcola $F(t) = P(X \le t) = \int_{-\infty}^{t} f(x)\,dx$ con $f(x)$ densità
- `qnorm(α, μ, σ)` Calcola il quantile di ordine α per una distribuzione $X \sim N(\mu, \sigma^2)$ cioè



- `dnorm(x, μ, σ)` Restituisce il valore numerico della densità e non il disegno
  Data $X \sim N(\mu, \sigma^2)$



`dnorm(x₀, μ, σ)` = $f(x_0)$

- `rnorm(n, μ, σ)` Genera n numeri gaussiani a caso
  Es. `x = rnorm(100, 0, 1)` . Li calcola e li assegna ad x (che quindi è un vettore)

  `hist(x, 10)` disegna un istogramma dei numeri assegnati a x, suddivisi in 10 intervalli

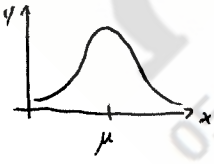
③ Come tracciare il grafico di una densità gaussiana con R ?

Il comando dnorm restituisce solo il valore numerico della densità, ma non il disegno. Per costruire il grafico devo determinare coppie di x e y come segue:

`x = (0:1000)/1000 * 10 * σ - 5 * σ` , cioè mille-uno numeri compresi tra -5σ e 5σ (max)

`y = dnorm(x, μ, σ)`

`plot(x, y)`



esempio `x = (0:1000)/1000 . 6 - 3`
`y = dnorm(x, 0, 1)`
`plot(x, y)`

④ Come tracciare un istogramma relativo a dati sperimentali ?

`x = rnorm(100, 0, 1)`

`hist(x, 10)` Disegna un istogramma dei numeri assegnati a x, suddiviso in 10 intervalli

Master Copisteria 050/8312124 388/9837743

Covarianza, correlazione e formule empiriche

⑤ Definizione di covarianza

Date X, Y v.a., si ha che $\text{cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$ con $\mu_X = E[X]$ e $\mu_Y = E[Y]$

N.B.
- $\text{cov}(X,X) = \text{var}(X)$
- la covarianza è lineare in entrambi gli argomenti quindi

$$\text{cov}(\alpha X_1 + \beta X_2 + \gamma, Y) = \alpha\,\text{cov}(X_1,Y) + \beta\,\text{cov}(X_2,Y) + \gamma\,\text{cov}(1,Y) =$$
$$= \alpha\,\text{cov}(X_1,Y) + \beta\,\text{cov}(X_2,Y) \quad \text{con } \alpha,\beta,\gamma \in \mathbb{R}$$

la stessa cosa vale per l'argomento Y

Ricordiamo: Teorema 1 $\quad \text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{cov}(X,Y)$

Teorema 2 Se X, Y indipendenti, allora $\text{cov}(X,Y) = 0$ ~~⇔~~

⑥ Invarianza di scala (o meno) della covarianza

Come la varianza, che si comporta come un quadrato, anche la covarianza soffre di problemi di scala. Per la varianza si estrae la radice quadrata introducendo la deviazione standard; mentre per la covarianza, che non è sempre positiva, si effettua una normalizzazione dividendo per le deviazioni standard (si ottiene ρ).

⑤ Definizione di correlazione

Si chiama coefficiente di correlazione tra X e Y il numero definito da

$$\text{corr}(X,Y) = \rho(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X\,\sigma_Y} \quad \text{con } \sigma_X^2 = \text{Var}(X),\ \sigma_Y^2 = \text{Var}(Y)$$

(è la normalizzazione accennata sopra)

Teorema 1 $\quad -1 \le \rho(X,Y) \le 1$

Teorema 2 Se X, Y indipendenti, allora $\rho(X,Y) = 0$, dato che $\text{cov}(X,Y) = 0$ ~~⇔~~

⑥ Invarianza di scala (o meno) della correlazione

La correlazione non risente di un cambio di scala (di unità di misura) infatti,

$$\rho(\lambda X, \lambda Y) = \rho(X,Y) \quad \text{con } \lambda = \text{parametro numerico}$$

⑦ Legami tra indipendenza e condizioni su varianza e correlazione

Se X e Y sono v.a. indipendenti allora $\text{cov}(X,Y) = 0$ e $\rho(X,Y) = 0$

Viceversa se $\text{cov}(X,Y) = 0$, non è detto che X e Y siano indipendenti.

Diciamo che X e Y sono scorrelate se hanno correlazione nulla. Quindi l'indipendenza implica la scorrelazione.

MASTER COPY
Tel. 388/9837745

⑤ Definizione di covarianza empirica (sperimentale) e correlazione empirica (sperimentale) ④

$$\hat{cov} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \qquad \text{con} \quad \bar{x}=\frac{x_1+\dots+x_n}{n}, \quad \bar{y}=\frac{y_1+\dots+y_n}{n}$$

avendo n coppie di valori sperimentali $(x_1,y_1)\dots(x_n,y_n)$ cioè una tabella in cui le colonne corrispondono alle variabili e le righe agli individui (unità sperimentali, unità osservate)

	X	Y
1	x_1	y_1
⋮		
n	x_n	y_n

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Questi indicatori sono buone stime di quelli teorici, essendo applicate le formule teoriche a dati sperimentali

⑧ Vicinanza a zero di $\hat{\rho}$ rispetto alla vicinanza a zero di $\hat{cov}$

La vicinanza a zero della correlazione empirica ($\hat{\rho}\simeq 0$) si interpreta come sintomo di indipendenza o comunque come bassa dipendenza, la vicinanza ad 1 come elevato legame positivo, la vicinanza a -1 come elevato legame negativo. Invece, dipendendo il numero di $\hat{cov}$ dalla scala scelta, la sua vicinanza a zero è meno assoluta, quindi può trarre in inganno

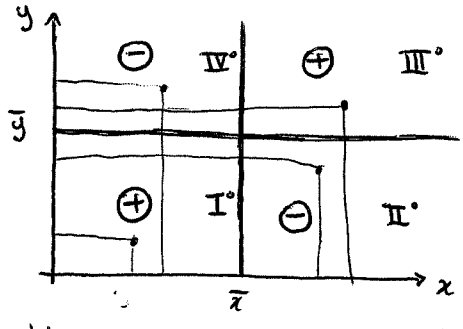
⑨ Come è fatta una tabella di dati sperimentali relativi ad una coppia di variabili (X,Y)

Stiamo pensando di avere n coppie di valori sperimentali $(x_1,y_1)\dots(x_n,y_n)$ o più espressamente una tabella in cui le colonne sono le variabili e le righe gli individui (o unità sperimentali, unità osservate)

	X	Y
1	x_1	y_1
⋮		
n	x_n	y_n

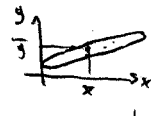
MASTER COPY
Tel. 388/9837745

⑩ ⑪ Interpretazione grafica del segno di $\hat{cov}$, sulla base della formula empirica

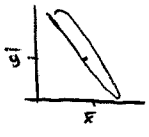


Sulla base della formula empirica
- nella I° zona del grafico $(x_i-\bar{x})<0$ e $(y_i-\bar{y})<0$ quindi $(x_i-\bar{x})(y_i-\bar{y})>0$, cioè $\hat{cov}$ ha segno positivo
- nella II° zona, $\hat{cov}$ ha segno negativo
- nella III° zona, $\hat{cov}$ ha segno positivo
- nella IV° zona, $\hat{cov}$ ha segno negativo

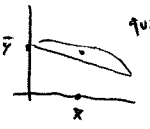
Se i dati sperimentali formano una nuvola di punti di questo tipo



quasi orizzontale $\hat{cov}\simeq 0$ (e maggiore di zero) e $\hat{\rho}\simeq 1$



$\hat{cov}\simeq -1$
$\hat{\rho}\simeq -1$



quasi orizzontale
$\hat{cov}\simeq 0$ (negativo)
$\hat{\rho}\simeq -1$



$\hat{cov}\simeq 0$
$\hat{\rho}\simeq 0$

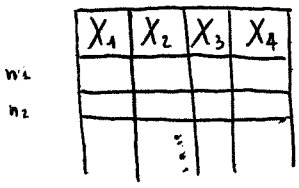
Il segno di $\hat{\rho}$ dipende dal segno di $\hat{cov}$.

Vettori Aleatori

Definizione: Una sequenza finita $X=(X_1, \dots, X_n)$ di v.a. è un vettore aleatorio

Vettore dei valori medi: $m = (m_1, \dots, m_n)$ con $m_k = E[X_k]$

Matrice di covarianza: La matrice di covarianza Q di un vettore $X=(X_1, \dots, X_n)$ è una matrice i cui elementi sono $Q_{ij} = \mathrm{cov}(X_i, X_j)$



Con la Q si guarda se le X_i sono legate tra di loro (sempre a due a due).

(11) Proprietà della matrice di covarianza di un vettore aleatorio (solo enunciato)

La matrice di covarianza Q di un vettore aleatorio $X=(X_1 \dots X_n)$ è una matrice i cui elementi sono $Q_{ij}=\mathrm{cov}(X_i,X_j)$

È una matrice simmetrica, semi-definita positiva e presenta sulla diagonale principale le varianze $Q_{ii}=\mathrm{Var}[X_i]$

Per quanto riguarda le trasformazioni lineari se ho $X=(X_1, \dots, X_n)$ e $Y=(Y_1, \dots, Y_m)$, la trasformazione $Y = AX+b$ (con $A \in \mathbb{R}^{n \times m}$ e $b \in \mathbb{R}^m$) è tale che $Q_Y = A Q_X A^T$ e $\mu_Y = A\mu_X + b$

(12) Proprietà della matrice di covarianza e del vettore dei valori medi di un vettore aleatorio, anche sotto trasformazioni lineari (con dimostrazioni)

- È una matrice simmetrica, cioè $Q^T = Q$

  dim $Q_{ij} = \mathrm{cov}(X_i, X_j) = \mathrm{cov}(X_j, X_i) = Q_{ji}$

- Sulla diagonale ci sono le varianze

  dim $Q_{ii} = \mathrm{cov}(X_i, X_i) = \mathrm{var}(X_i)$

- È una matrice semi-definita positiva cioè $v^T \cdot Q \cdot v \geq 0 \quad \forall v$ (v^T)(Q)(v) = un numero

  dim
  $$\sum_{i,j} v_i Q_{ij} v_j = \sum_{i,j} v_i \cdot \mathrm{cov}(X_i, X_j) \cdot v_j = \sum_{i,j} \mathrm{cov}(v_i X_i, v_j X_j) = \mathrm{cov}\left(\sum_i v_i X_i, \sum_j v_j X_j\right) =$$

  ↑ linearità nei due argomenti della covarianza

  ↑ linearità rispetto alla sommatoria

  ↑ i e j sono uguali perché Q è simmetrica, sono quindi indici muti.

  $$= \mathrm{cov}(Y, Y) = \mathrm{var}[Y] \geq 0$$

  se fosse >0 sarebbe definita positiva