

INTRODUZIONE AL MACHINE LEARNING

F. Morandin

ord 1

18/02/2025

Note Title

• orario : mar 13:30 - 16:30 teoria → 32 h

mer 8:30 - 10:30 laboratorio → 24 h

LM sc. inf

ricoveramento : mer 15:30

LM mat

• Cose pratiche

prerequisiti : elementi di probabilità Cap 5 - 8 Ross + algebra lineare + calcolo

Elly : INTROD. AL MACHINE LEARNING

Telegram

Biblio : note 2024 + Ross + James - Hastie - ... <https://www.statlearning.com/>

Esame : prova pratica su python ~ 5 h + orale

→ allo scritto si possono usare appunti cartacei

→ quest'anno laboratori su python (da zero)

→ allo scritto ci sarà accesso a internet

→ video su youtube

Motivazione e contenuti

storia : 2005 ing. gestionale 9 cfu : EDA. + int ML pratico

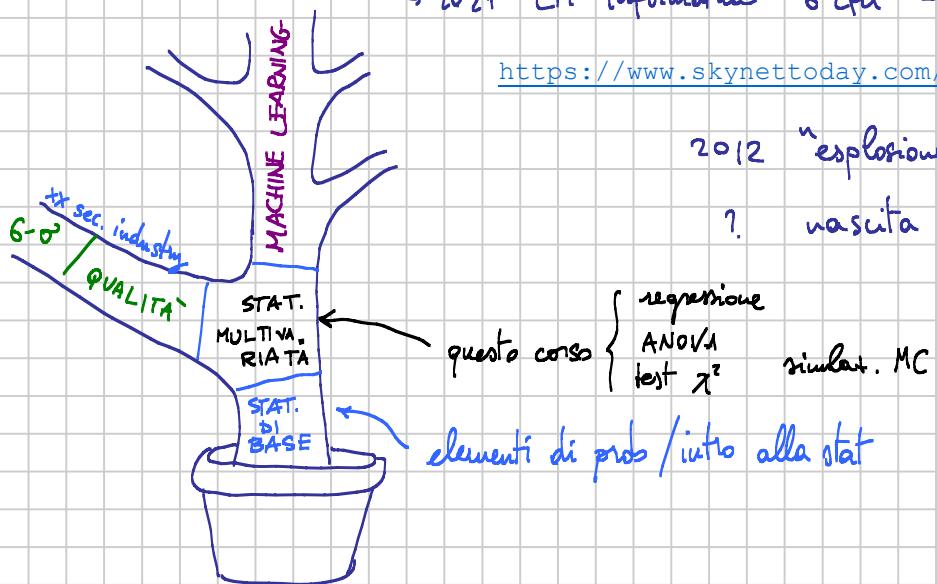
↳ 2016 LM matematica 6 cfu + teoria molti meno contenuti ST. Ind.

↳ 2021 LM informatica 6 cfu - teoria + verso ML (in direzione)

<https://www.skynettoday.com/overviews/neural-net-history>

2012 "esplosione" del Deep Learning

? nascita Big Data

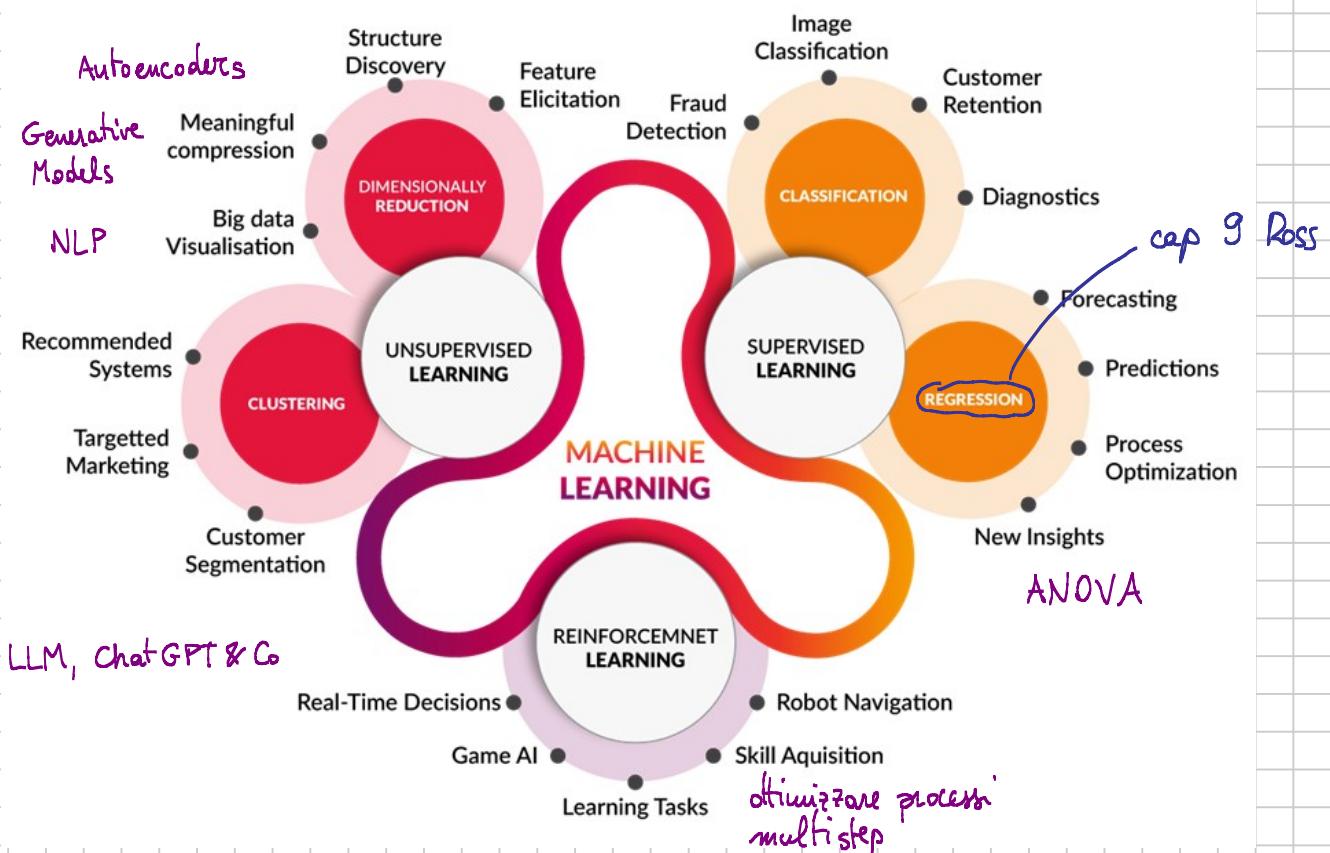


cosa vi servirà nella vita? ↗ statistiche in azienda
↗ machine learning /d.l.

- quale software?

- Excel : veloce nell'uso , risultati parziali visibili , più accettato dalle aziende
- python : moderno , efficace , elegante , obbligato per le reti neurali
- R : bioinformatica , molti pacchetti per data science , grafici bellissimi
- minitab : statistica avanzata aziendale

- cos'è il machine learning



Documentario : Alpha Go (yt)

<https://youtu.be/wXuK6gekU1Y>

<http://frmor.net/>

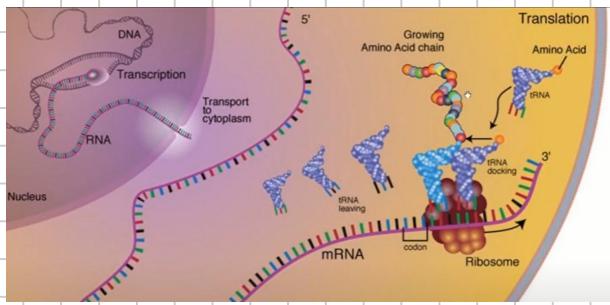
<https://github.com/sai-dev/sai#what>



<https://youtu.be/nRb1z6GKWvI>

Deep learning methods: a practical approach
FRANCESCO MORANDIN

https://youtu.be/F_ykpgIO2A8



possibilità di tenere su Deep Learning, Reinforcement Learning, Bioinformatica

- Cos'è un dataset (tipico, e in forma standard)

<https://archive.ics.uci.edu/>

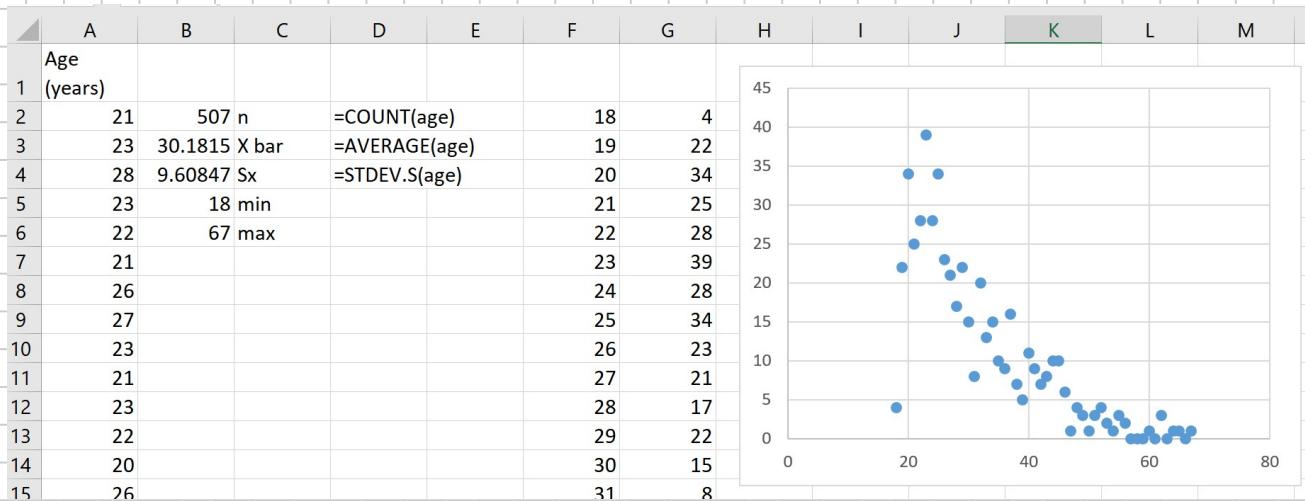
<https://www.kaggle.com/>

è una tabella che ha:

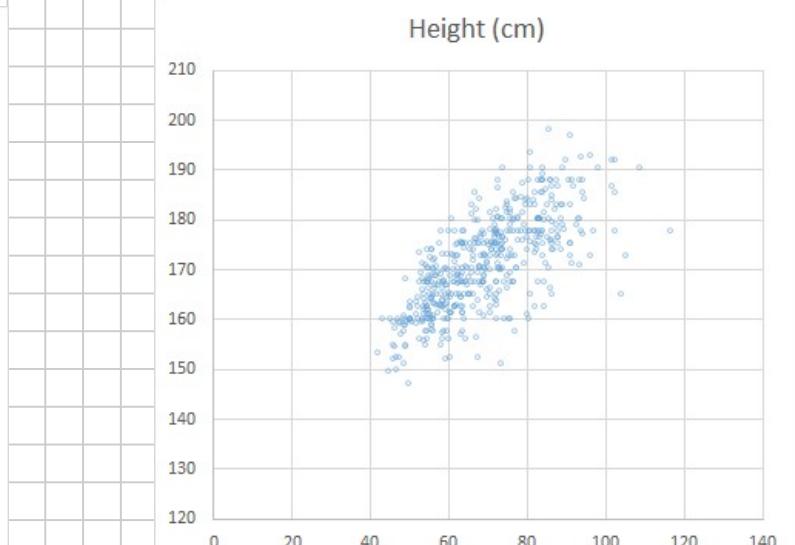
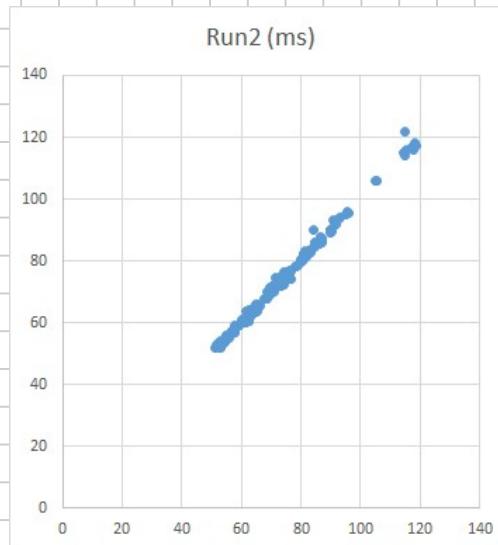
righe = record = item

colonne = field = variabili

→ ogni variabile può essere considerata un campione statistico e studiata come tale



→ più variabili insieme hanno distribuzioni congiunte che potrebbero studiare



RIPASSO PROPRIETÀ VV. AA.

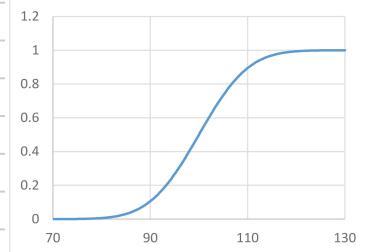
ora 2

• Cdf / pdf / pmf

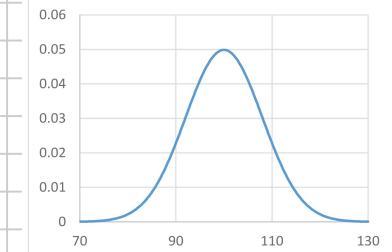
* v.a. continue hanno

Cdf F TRUE
e pdf f FALSE

=NORM.DIST(A4, μ , σ , TRUE)



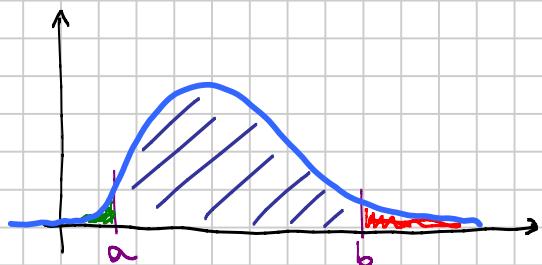
=NORM.DIST(A4, μ , σ , FALSE)



$$f = F'$$

$$F_x(t) = P(X \leq t)$$

* occhio alle funzioni Excel !!



▀	coda sx	$F_x(a)$
▀	coda dx	$1 - F_x(b)$
▢	intervallo	$F_x(b) - F_x(a)$

* v.a. discrete hanno

Cdf F TRUE
e pmf φ FALSE

$$F_x(k) = P(X \leq k)$$

$$\varphi_x(k) = F_x(k) - F_x(k-1)$$

$$F_x(k) = \sum_{j=0}^k \varphi_x(j)$$



▀	coda sx	$F_x(a)$
▀	coda dx	$1 - F_x(b-1)$
▢	intervallo (estri.esclusi)	$F_x(b-1) - F_x(a)$

• Altri usi delle CDF

a) Invertire le formule

ad es.: chiedo x : $P(\text{gamma}(\alpha, \beta) < x) \geq 5\%$

$$0.05 \leq P(\text{gamma}(\alpha, \beta) \leq x) = F_{\text{gamma}(\alpha, \beta)}(x)$$

per invertire, applico $F_{\text{gamma}(\alpha, \beta)}^{-1}$ a entrambi i membri

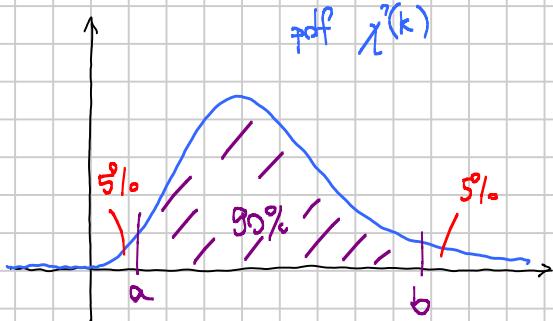
$$F_g^{-1}(0.05) \leq F_g^{-1}(F_g(x)) = x$$

perché F e F^{-1} sono sempre crescenti

Concludo che vanno bene tutti gli $x \geq \boxed{\text{GAMMA.INV}(0.05, \alpha, \beta)}$

b) Invertire le formule caso particolare: un dato e due incognite

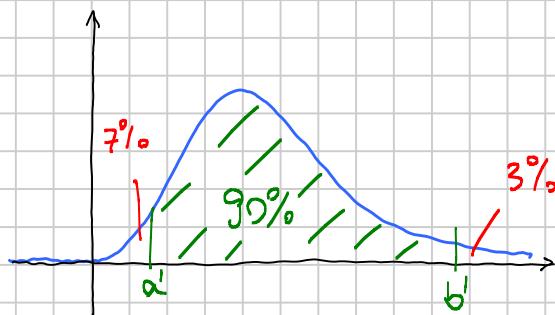
ad es.: chiedo $a \leftarrow b$: $P(\chi^2(k) \in [a, b]) = 90\%$



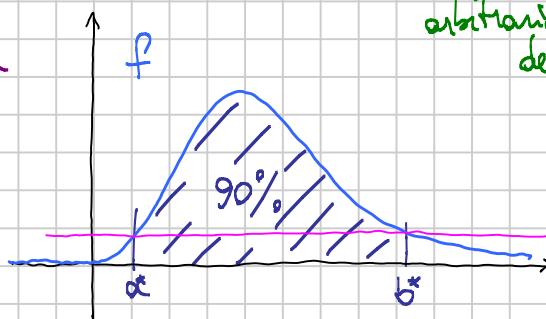
casuico: stessa prob per le due code

$$a = F^{-1}(5\%)$$

$$b = F^{-1}(95\%)$$



arbitrario: basta che la somma delle code sia quella richiesta



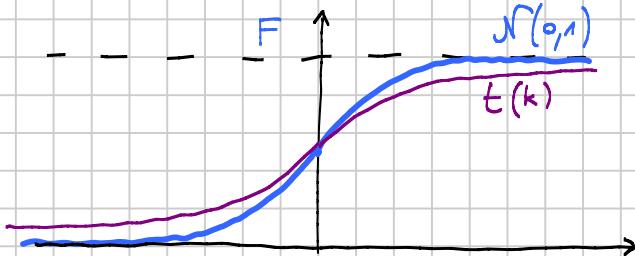
minimale: $f(a^*) = f(b^*)$ e somma delle code solita

\hookrightarrow intervallo di larghezza minima

* se la distribuzione è simmetrica, casuico = minimale

Hw: fare su Excel per $k=10$

c) Sulla simmetria delle CdF



$$\Phi(-x) = 1 - \Phi(x)$$

$$\Phi^{-1}(x) = -\Phi^{-1}(1-x)$$

$$\Phi(-x) + \Phi(x) = 1$$

* HW : Vedere tutte le funzioni corrispondenti su Excel

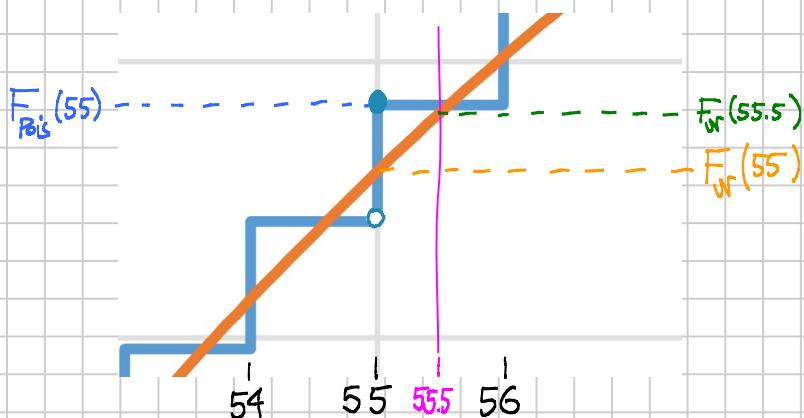
* Recall : correzione della continuità

X v.a. a valori interi ; F_x non disponibile

ma $F \approx F_x$ disponibile , F continua

$$F_x(k) = P(X \leq k) \approx F(k + \frac{1}{2}) \text{ più preciso di } F(k)$$

↑ ↑
 intero c.c.



d) Generazione di v.v.a.a. di legge data

→ distribuzione fissata (ad es. expo(λ)) → ottenga la CdF

i. cerco la CdF corrispondente $F(t) = 1 - e^{-\lambda t}$, $t \geq 0$

ii. calcolo l'inversa $y = 1 - e^{-\lambda t} \Leftrightarrow 1-y = e^{-\lambda t} \Leftrightarrow -\log(1-y) = \lambda t$
 $\Rightarrow t = -\frac{1}{\lambda} \log(1-y)$

$$F^{-1}(y) = -\frac{1}{\lambda} \log(1-y)$$

iii. compongo l'inversa con rand() che genera uniformi su $(0,1)$

$$X = F^{-1}(\text{rand}())$$

0.12 λ

$$3.30161 = -1/\lambda * \ln(1-\text{RAND}())$$

\rightarrow sia $U \sim \text{unif}(0,1)$ allora $F^{-1}(U)$ ha la legge data
 $\text{gamma}(\alpha=1) \sim \text{expo}$
 $= \text{GAMMA.INV}(\text{RAND}(), 1, 1/\lambda)$ e' expo (λ)

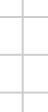
\rightarrow verifica :

$$X := F^{-1}(U)$$

voglio verificare se $F_x(t) = F(t)$

$$F_x(t) := P(X \leq t) = P(F^{-1}(U) \leq t) = P(U \leq F(t)) = F_U(F(t)) = F(t)$$

F è crescente



$$F_U(x) = \begin{cases} x & 0 < x < 1 \\ 0 & \\ 1 & \end{cases}$$

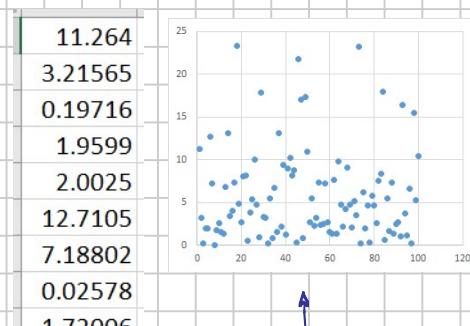


H.W.: $-\ln(\text{RAND}()) / \lambda$ è expo (λ)

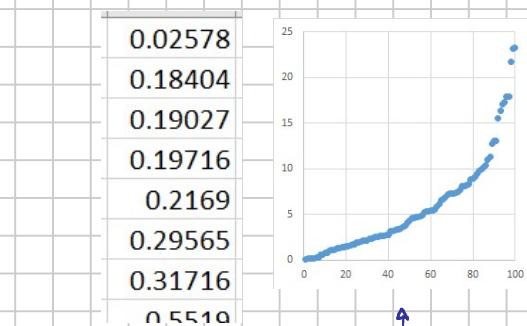
e) CdF empirica e diagramma Q-Q

Suppongo di avere un campione proveniente da una distribuzione con CdF F

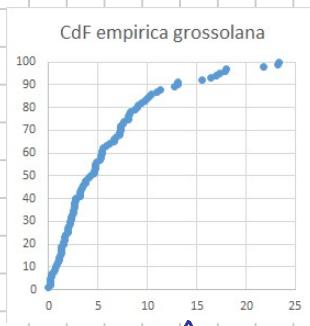
$x_1, x_2, \dots, x_n \longrightarrow x_{(1)}, x_{(2)}, \dots, x_{(n)}$ ordinato



(i, x_i) $i=1, 2, \dots, 100$

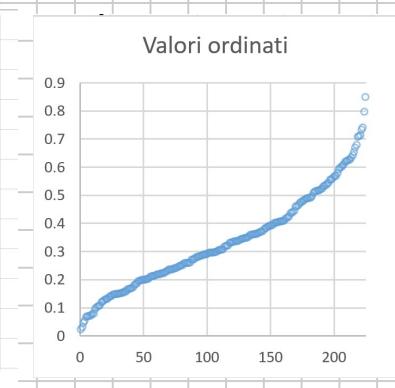
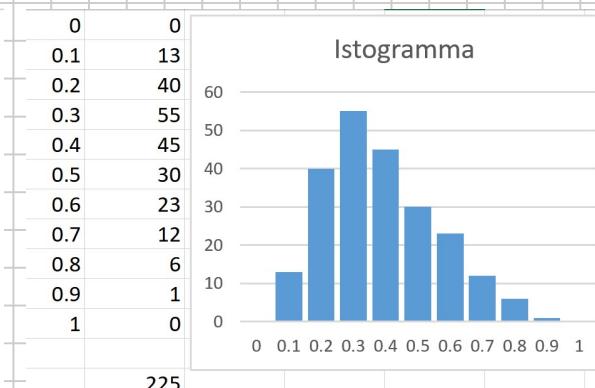
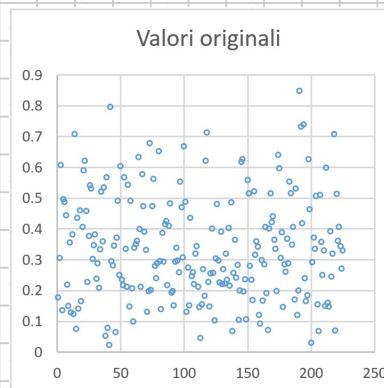


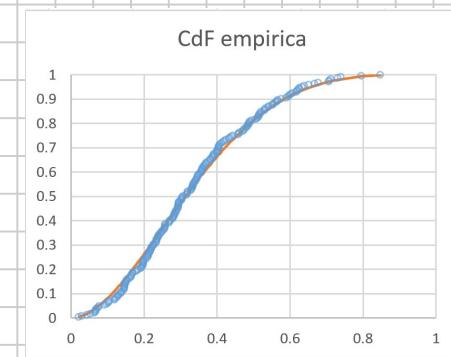
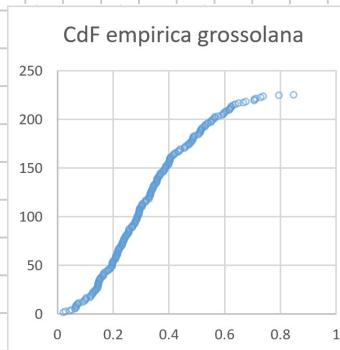
$(i, x_{(i)})$



$(x_{(i)}, i)$

* Legge beta (2, 4)





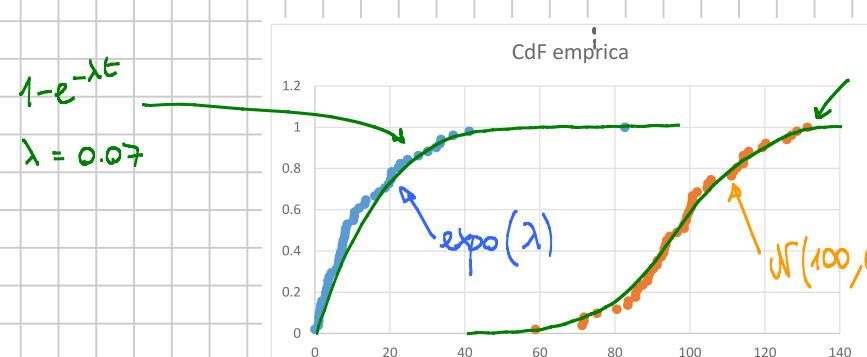
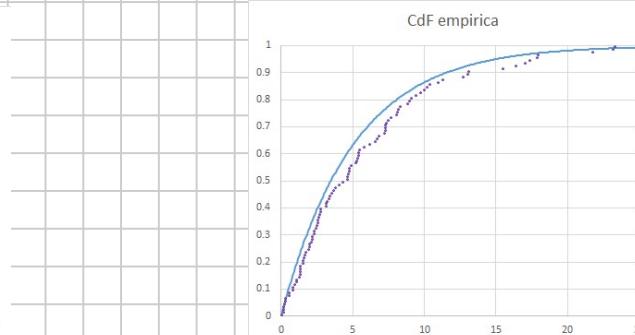
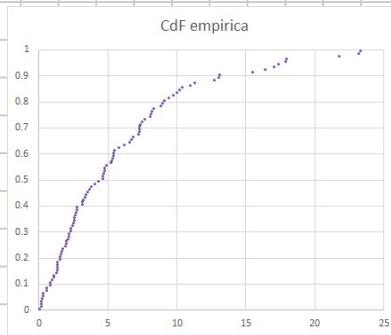
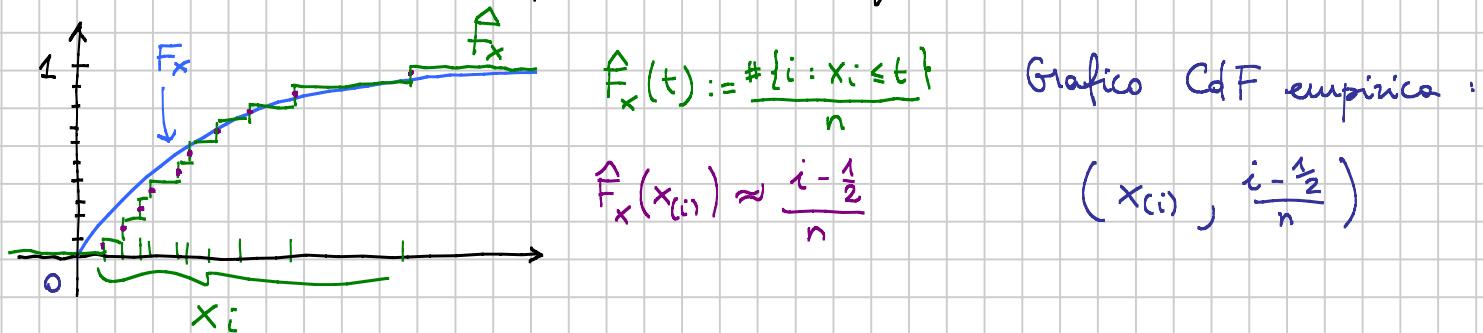
X Y

1	$(X_1 - 0.5) / 225$
2	0.00667

* cos'è la CdF empirica?

$$F_X(t) := P(X \leq t) \approx \frac{\#\{i : x_i \leq t\}}{n} =: \hat{F}(t) \quad \text{CdF empirica}$$

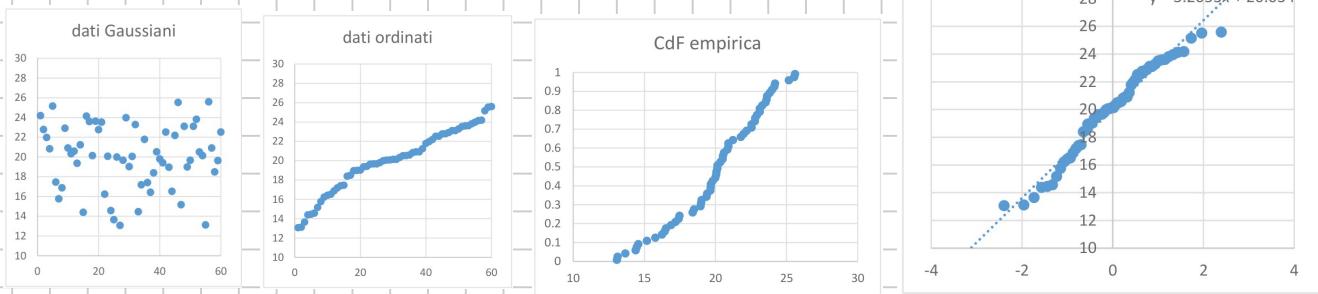
→ oss: $\hat{F}(x_{(i)}) = \frac{i}{n}$ quindi i punti $(x_{(i)}, \frac{i-0.5}{n})$ $i=1, \dots, n$
ci danno una approssimazione del grafico di F



→ diagramma Q-Q

↪ viene una retta in caso di dati Gaussiani
plotto i punti $(\Phi^{-1}(\frac{i-0.5}{n}), x_{(i)})$

$$\Phi = F_{N(\mu, \sigma^2)}$$



Qual è la logica?

Suppongo che $x_i \sim N(\mu, \sigma^2)$ e considero le due trasformazioni $F_{N(\mu, \sigma^2)}^{-1}$ e $F_{N(\mu, \sigma^2)}^{-1}$
che cambiano $N(\mu, \sigma^2)$ in $\text{unif}(0,1)$ e viceversa:

$$\begin{cases} p_i := F_{N(\mu, \sigma^2)}^{-1}(x_i) \\ x_i = F_{N(\mu, \sigma^2)}^{-1}(p_i) \end{cases} \quad x_i \sim N(\mu, \sigma^2) \Leftrightarrow p_i \sim \text{unif}(0,1)$$

Osserviamo anche che $F_{N(\mu, \sigma^2)}^{-1}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ quindi $F_{N(\mu, \sigma^2)}^{-1}(p) = \mu + \sigma \Phi^{-1}(p)$

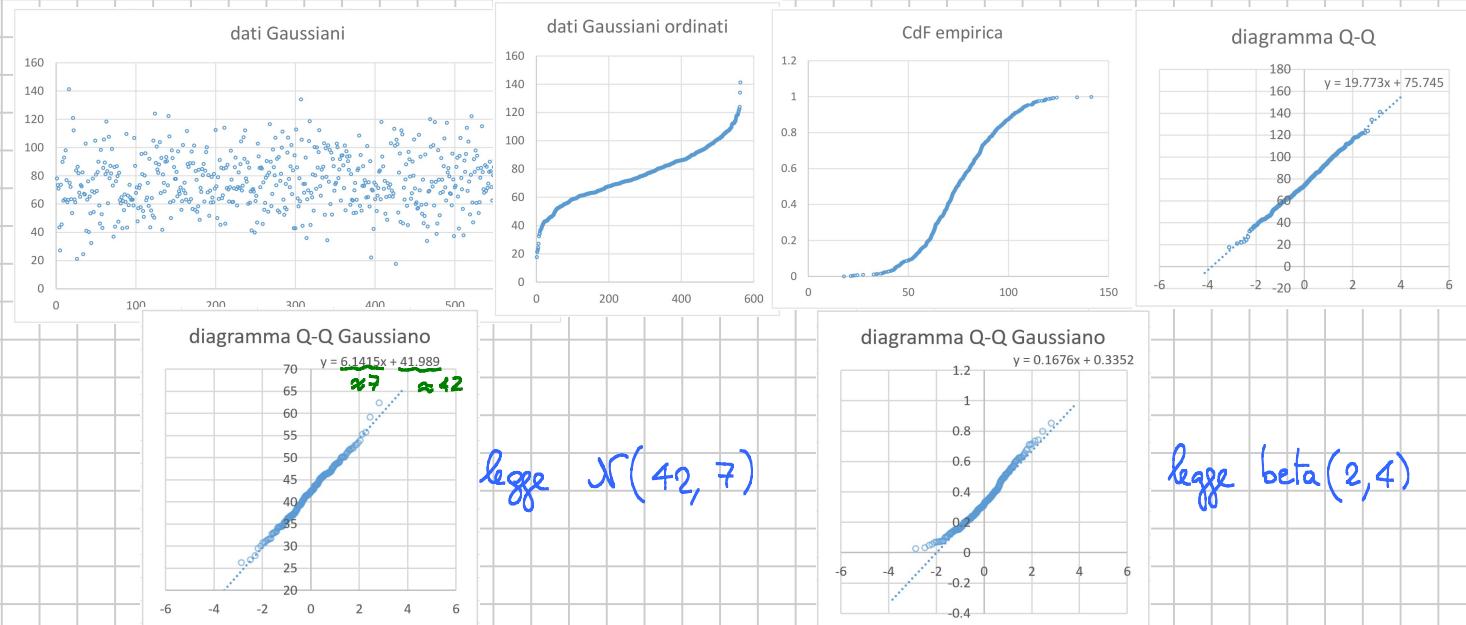
$$p_i = \Phi\left(\frac{x_i - \mu}{\sigma}\right) \Leftrightarrow \Phi^{-1}(p_i) = \frac{x_i - \mu}{\sigma} \Leftrightarrow x_i = \mu + \sigma \Phi^{-1}(p_i)$$

e quindi $x_{(i)}$, quindi anche $p_{(i)}$: il diagramma Q-Q diventa

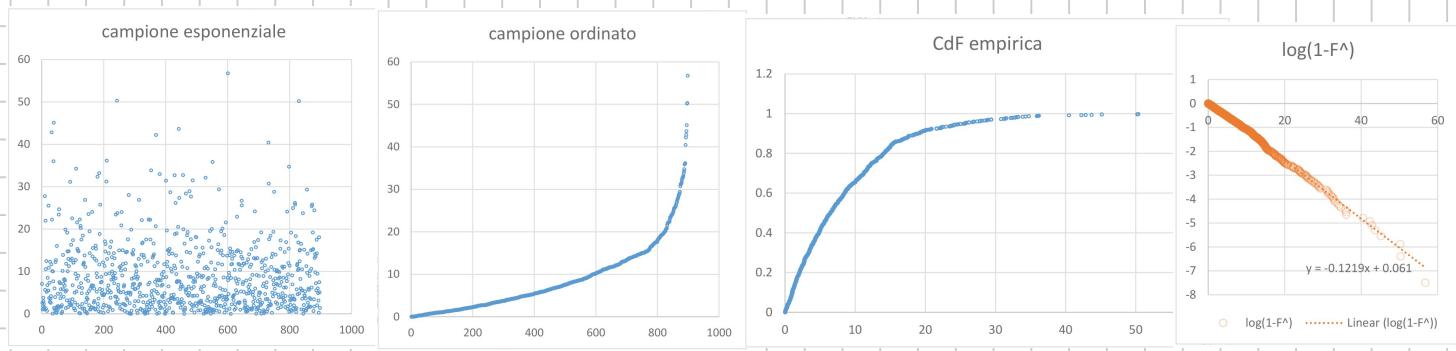
$$\text{QQ: } (\Phi^{-1}\left(\frac{i-0.5}{n}\right), \mu + \sigma \Phi^{-1}(p_{(i)})) \quad \text{chiamiamo } z_i = \Phi^{-1}\left(\frac{i-0.5}{n}\right)$$

$$\text{Inoltre } p_i \sim \text{unif}(0,1) \Rightarrow p_{(i)} \approx \frac{i-0.5}{n} \Rightarrow \Phi^{-1}(p_{(i)}) \approx z_i$$

$$\Rightarrow \text{QQ} \approx (z_i, \mu + \sigma z_i) \leftarrow \text{punti sulla retta } y = \mu + \sigma x$$



* Vale qualcosa di analogo anche nel caso esponenziale



LEGGI DI VVAI IMPORTANTI

GAUSSIANA

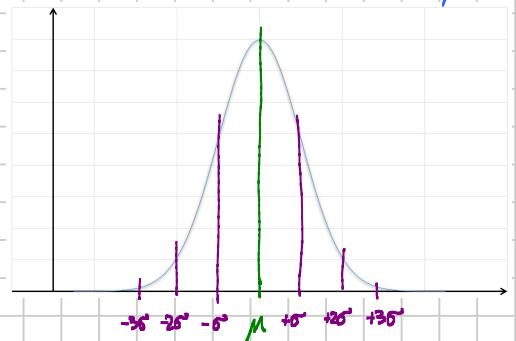
$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mu \in \mathbb{R} \quad \sigma^2 > 0$$

Quando $\sigma^2 > 0$ è continua con pdf

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R}$$

estendiamo al caso limite della "delta di Dirac" in μ



→ TLC: grandezze casuali che sono somma di tanti piccoli contributi indipendenti hanno legge $\sim \mathcal{N}$

→ classe chiusa per trasformazioni lineari

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad a+bX \sim \mathcal{N}(\text{opportuni parametri})$$

* classe = forma

→ riproducibile (super! non richiede X, Y indipendenti!)

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad Y \sim \mathcal{N}(\nu, \tau^2) \quad X+Y \sim \mathcal{N}(\mu+\nu, \text{dipende})$$

→ CdF canonica: $\Phi(t) = F_{\mathcal{N}(0,1)}(t) = P(N(0,1) \leq t)$

$$F_{\mathcal{N}(\mu, \sigma^2)}(x) = P(N(\mu, \sigma^2) \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

LOGNORMALE

$$X \sim \text{lognorm}(\mu, \sigma^2) \quad \mu \in \mathbb{R} \quad \sigma > 0$$

si tratta dell'esponentiale di una Gaussiana : è positiva

→ il suo logaritmo è una v.a. normale $\log X \sim \mathcal{N}(\mu, \sigma^2)$

→ cambiare la base del logaritmo è solo una trasformazione lineare
della Gaussiana naturale

$$X \sim \text{lognorm}(\mu, \sigma^2) \quad \Leftrightarrow \log X \sim \mathcal{N}(\mu, \sigma^2)$$

$$\log_{10} X = \frac{\log X}{\log 10} \sim \mathcal{N}\left(\frac{\mu}{\log 10}, \left(\frac{\sigma}{\log 10}\right)^2\right)$$

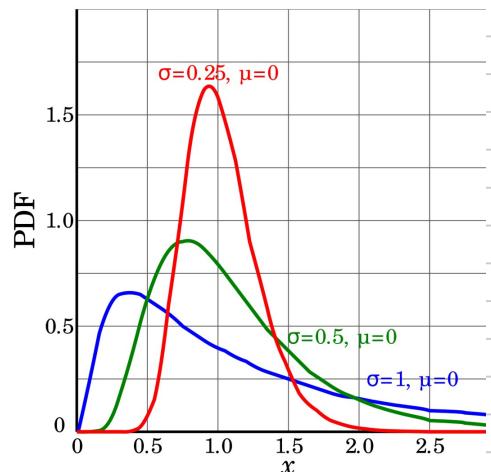
→ è asimmetrica, con coda a dx

→ assume spesso valori con diversi ordini di grandezza

→ analisi di dati : si fa il logaritmo e si trattano dati normali

$$X \sim \text{lognorm}(\mu, \sigma^2) \quad E(X) = e^{\mu + \frac{\sigma^2}{2}}$$

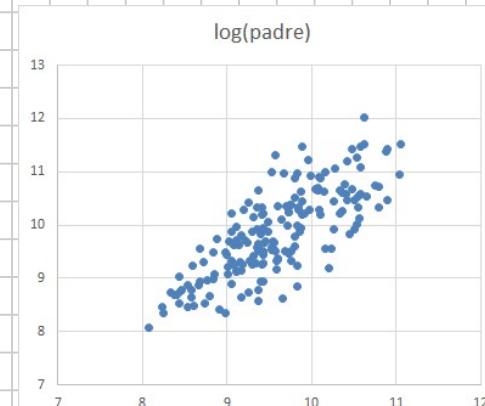
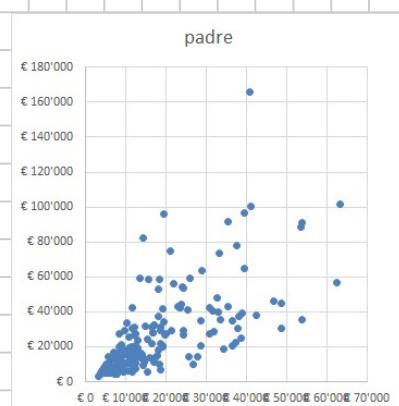
HW: la mediana è e^μ



→ TLC : grandezze normali che sono il prodotto di tanti piccoli contributi
indipendenti hanno legge $\sim \text{lognorm}$

* ad esempio : redditi, patrimoni, dimensioni frammenti

patrimonio tipico :	100k	$\begin{cases} \text{medio} & +50\% \\ \text{famiglia povera} & -70\% \\ < 30 & -80\% \end{cases}$	$\times 1.5$
			$\times 0.3$
			$\times 0.2$



GAMMA | CHI-QUADRO | ERLANG
(anche expo è un caso particolare)

$$X \sim \text{gamma}(\alpha, \lambda) \quad X \sim \text{gamma}(\alpha, \beta)$$

detto β su `scipy.stats`
 $\alpha > 0$
 $\lambda / \beta > 0$

Excel

ci sono due parametrizzazioni incompatibili

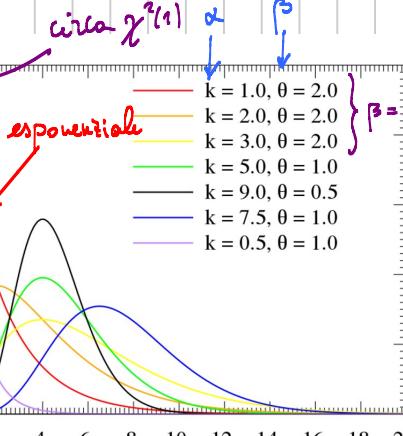
$$f_X(t) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha t^{\alpha-1} e^{-\lambda t}, \quad t > 0$$

$$= \frac{1}{\Gamma(\alpha)} \beta^{-\alpha} t^{\alpha-1} e^{-t/\beta}, \quad t > 0$$

- la forma è determinata da α
- la scala è determinata da $\lambda \circ \beta$

→ Se $\alpha = 1$ è esponentiale

🚩 $\text{gamma}(1, \lambda) \sim \text{expo}(\lambda)$



esattamente uguale

- Se $\alpha = n \geq 1$ è intero, si chiama anche Erlang ed è la somma di n $\text{expo}(\lambda)$ indipendenti
- 🚩 $T_1, T_2, \dots, T_n \sim \text{expo}(\lambda)$ iid. $\rightarrow \sum_i^n T_i \sim \text{gamma}(n, \lambda)$

→ è riproducibile a λ fissato

$$X \sim \text{gamma}(\alpha_1, \lambda) \quad Y \sim \text{gamma}(\alpha_2, \lambda) \quad \text{indip.}$$

$$X + Y \sim \text{gamma}(\alpha_1 + \alpha_2, \lambda)$$

→ $X \sim \text{gamma}(\alpha, \lambda)$

$$E(X) = \frac{\alpha}{\lambda} = \alpha \beta$$

$$\text{Var}(X) = \frac{\alpha}{\lambda^2} = \alpha \beta^2$$

→ trasformazioni lineari :

$$X \sim \text{gamma}(\alpha, \lambda)$$

$$cX \sim \text{gamma}\left(\alpha, \frac{\lambda}{c}\right)$$

$$\lambda' = \frac{\lambda}{c} \quad \beta' = c\beta$$

$d + X$ non è più di tipo gamma

→ Se $\alpha = \frac{k}{2}$ e $\lambda = \frac{1}{2}$ si chiama anche chi-quadro con k g.d.l.

$$\chi^2(k) \sim \text{gamma}\left(\frac{k}{2}, \frac{1}{2}\right) \sim \text{gamma}\left(\frac{k}{2}, \beta = 2\right)$$

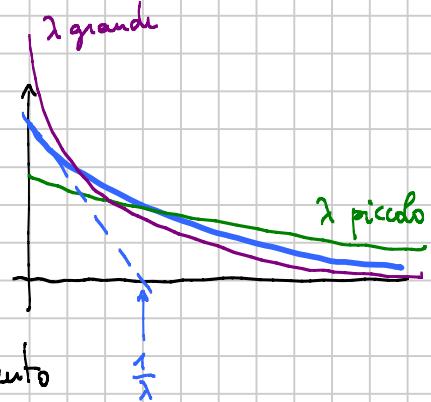
Hw / Recall : se $Z \sim \mathcal{N}(0,1)$, allora $Z^2 \sim \chi^2(1) \sim \text{gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$

• ESPOENZIALE

$T \sim \text{expo}(\lambda)$

$$f_T(t) = \lambda e^{-\lambda t}, \quad t \geq 0$$

"rate"
tasso - intensità



→ è la versione continua della geometrica

rappresenta un tempo di attesa di un evento

casiale che ha la stessa "probabilità" di avvenire in ogni momento

→ assenza di memoria : $P(T > a + b | T > a) = P(T > b)$

→ modellizza tempi di attesa per eventi improvvisi e imprevedibili come:
telefonate, notizie improvvise, decadimenti radioattivi

* CHI - QUADRO

$W \sim \chi^2(k) \quad k = 1, 2, 3, \dots$

$$\rightarrow W \sim \chi^2(k) \sim \text{gamma}\left(\frac{k}{2}, \frac{1}{2}\right) \quad E(W) = k \quad \text{Var}(W) = 2k$$

→ definizione operativa : $Z_1, Z_2, \dots, Z_k \sim \mathcal{N}(0,1)$ iid.

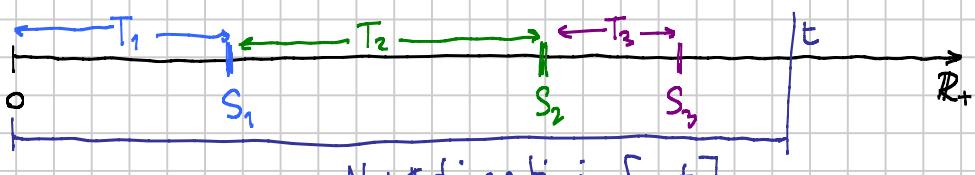
$$\text{allora } Z_1^2 + \dots + Z_k^2 \sim \chi^2(k)$$

$$\text{ciascuna } \chi^2(1) \sim \text{gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$$

* IL PROCESSO DI POISSON

→ Processo di Poisson : eventi istantanei con intervalli

$T_i \sim \text{expo}(\lambda)$ iid.



N_t : # di eventi in $[0, t]$

FLAG $T_i \sim \text{expo}(\lambda)$

FLAG $S_i \sim \text{gamma}(i, \lambda)$

FLAG $N_t \sim \text{Pois}(\lambda t)$

* TEORIA DELL'AFFIDABILITÀ

studia tempi di vita di cose → ruolo centrale dell'esponenziale

FLAG Cap 14 del Ross

INTRODUZIONE AL MACHINE LEARNING

ord 4

Note Title

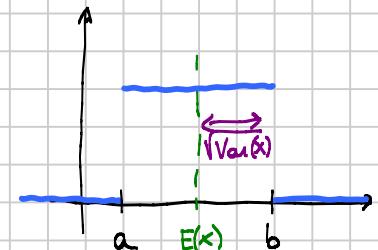
25/02/2025

UNIFORME

→ uniforme : $X \sim \text{unif}(a, b)$ $a < b$ reali

→ pdf $f_X(t) = c$, $a < t < b$

$$c = \frac{1}{b-a}$$



→ $\text{rand}() \sim \text{unif}(0, 1)$

→ $X \sim \text{unif}(a, b)$ $E(X) = \frac{a+b}{2}$ $\text{Var}(X) = \frac{(b-a)^2}{12}$

→ classe chiusa per trasformazioni lineari

BETA

$X \sim \text{beta}(\alpha, \beta)$

$\alpha, \beta > 0$

$$f_X(t) = C_{\alpha, \beta} \cdot t^{\alpha-1} \cdot (1-t)^{\beta-1}$$

→ $\text{beta}(1, 1) \sim \text{unif}(0, 1)$

→ $m, n \in \mathbb{N} \setminus \{0\}$

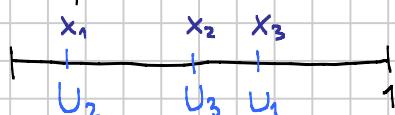
$\text{beta}(m, n)$ è la distribuzione

del m -esima più piccola

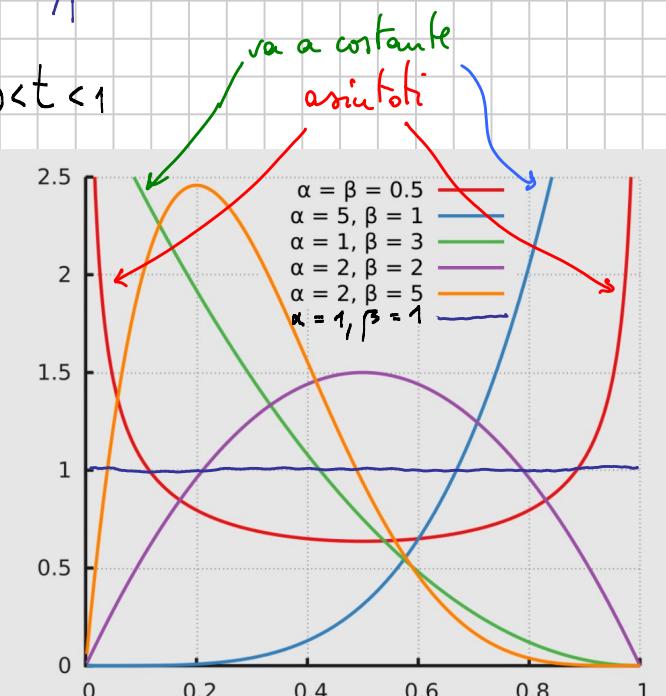
o n -esima più grande

tra $m+n-1$ $\text{unif}(0, 1)$ i.d.

$U_1, \dots, U_m \sim \text{unif}(0, 1)$ i.d.

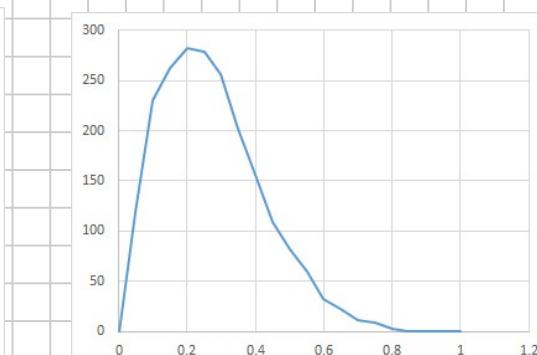
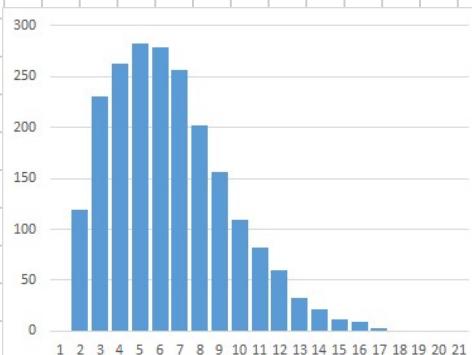


Le riordino : sia $X_i := U_{(i)}$



X_2 ha legge beta(2, 2) $X_1 \sim \text{beta}(1, 3)$

simulation con $m=2, n=5$ (vedi curva arancione)



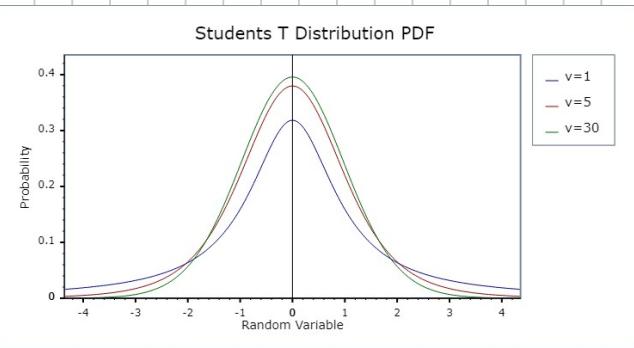
t DI STUDENT

$$X \sim t(k) \quad k=1, 2, \dots$$

→ pdf

$$f_x(t) = C_k \left(1 + \frac{1}{k} t^2\right)^{-\frac{k+1}{2}}$$

Hw: fare il limite per $k \rightarrow \infty$ con x fissato



$$\rightarrow t(k) \xrightarrow{k \rightarrow \infty} \mathcal{N}(0, 1) \quad E(x) = 0$$

→ definizione operativa: $Z \sim \mathcal{N}(0, 1)$, $W \sim \chi^2(k)$ indipendenti

🚩 $Z \cdot \left(\frac{W}{k}\right)^{\frac{1}{2}} \sim t(k)$

F DI FISHER

$$X \sim F(m; n) \quad m, n \in \{1, 2, 3, \dots\}$$

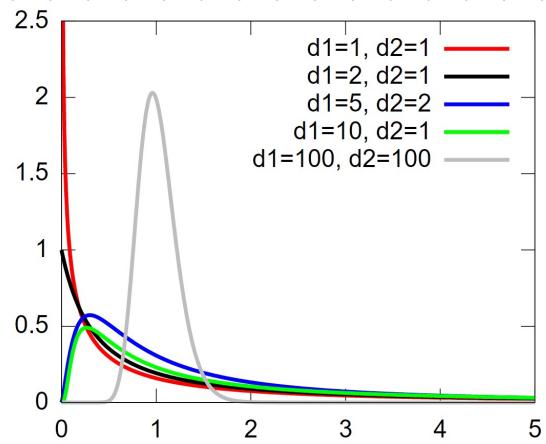
→ definizione operativa: $W_1 \sim \chi^2(m)$, $W_2 \sim \chi^2(n)$ indipendenti



$$\frac{W_1/m}{W_2/n} \sim F(m; n)$$

* si usa per confrontare due varianze campionarie

→ m, n sono i g.d.l. del numeratore e del denominatore



BINOMIALE

$$X \sim \text{bin}(n, p)$$

n intero positivo $p \in [0, 1]$

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k=0, 1, \dots, n$$

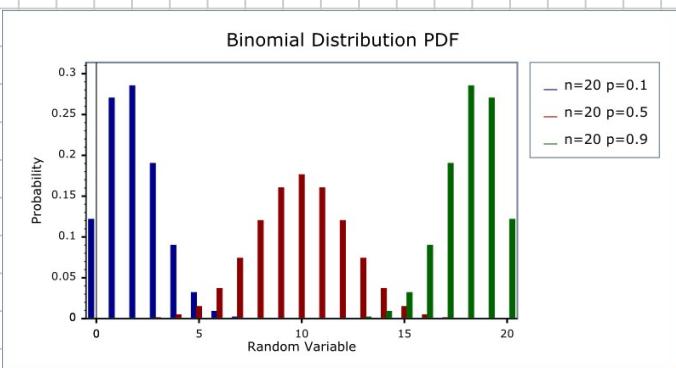
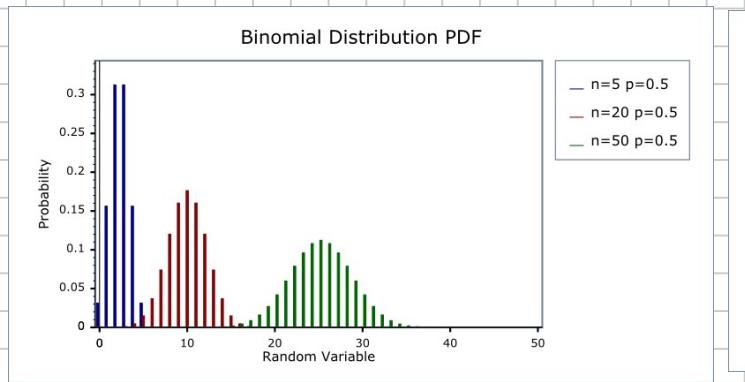
→ $n=1$ aka Bernoulliana

→ Se $\underbrace{Y_1, Y_2, \dots, Y_n}$ tutte $\text{bin}(1, p)$ i.i.d. allora $Y_1 + \dots + Y_n \sim \text{bin}(n, p)$

→ n prove indipendenti (sì/no) tutte con la stessa p , allora

il numero di prove con esito positivo ha legge $\text{bin}(n, p)$

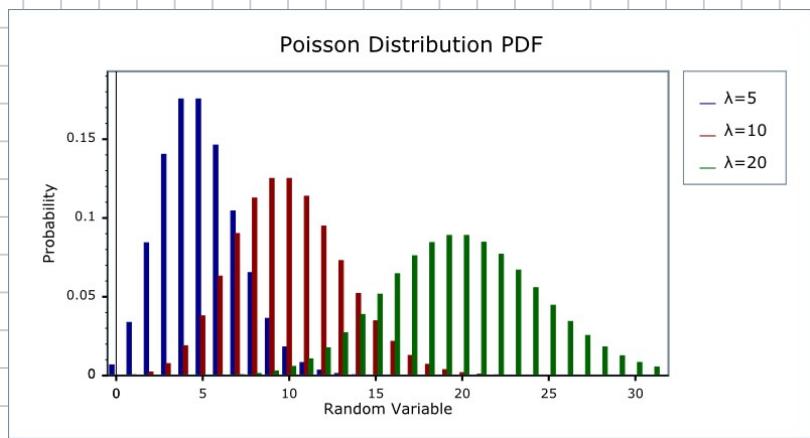
→ è riproducibile ovvero, se X_1, X_2, \dots, X_m indipendenti $X_i \sim \text{bin}(n_i; p)$ allora $X_1 + \dots + X_m \sim \text{bin}(n_1 + \dots + n_m, p)$



$$\rightarrow X \sim \text{bin}(n, p) \quad E(X) = np \quad \text{Var}(X) = np(1-p)$$

● **Poisson** $X \sim \text{Pois}(\nu)$ $\nu > 0$

$$P(X=k) = \frac{\nu^k}{k!} e^{-\nu} \quad k=0, 1, 2, \dots$$



quando ν è piccolo si vede che è asimmetrica

\rightarrow è il limite della binomiale per p piccolo e n grande

$$\text{bin}(n_k; p_k) \quad k=1, 2, \dots \quad \text{con } n_k \rightarrow \infty \quad p_k \rightarrow 0$$

$$\text{con } n_k p_k \rightarrow \nu > 0$$

$$\text{allora } \lim_{k \rightarrow \infty} \text{bin}(n_k; p_k) \xrightarrow{k \rightarrow \infty} \text{Pois}(\nu)$$

$$\lim_{k \rightarrow \infty} P(\text{bin}(n_k; p_k) = j) = P(\text{Pois}(\nu) = j)$$

* in pratica se $p \ll 1$ allora $\text{bin}(n, p) \approx \text{Pois}(np)$

HW: provare se fissa e "quanto" piccolo serve p (software)

\rightarrow è riproducibile: X_1, \dots, X_m indip. $X_i \sim \text{Pois}(\nu_i)$

$$\text{allora } X_1 + \dots + X_m \sim \text{Pois}(\nu_1 + \dots + \nu_m)$$

$$\rightarrow X \sim \text{Pois}(\nu) \quad E(X) = \nu \quad \text{Var}(X) = \nu \quad \text{flag}$$

→ conta il numero di successi in cari in cui bin non va bene
perché n e p sono più vaghi di np (la media)

- * numero di iscritti a informatica
- * numero di gol di una squadra in una partita

$n = 30$ minuti $p = 1\%$ prob gol per minuto

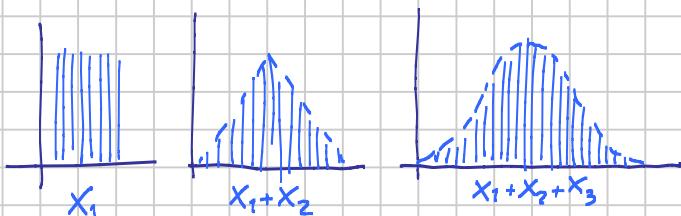
$n' = 30 \cdot 60$ secondi $p' \approx \frac{1}{6000}$ prob di gol per secondo

} $\sim \text{Pois}(0,9)$

• UNIFORME DISCRETA come il dado

$$P(X=i) = \frac{1}{n} \quad i=1, 2, \dots, n \quad \text{int}(\text{rand()} \cdot n) + 1$$

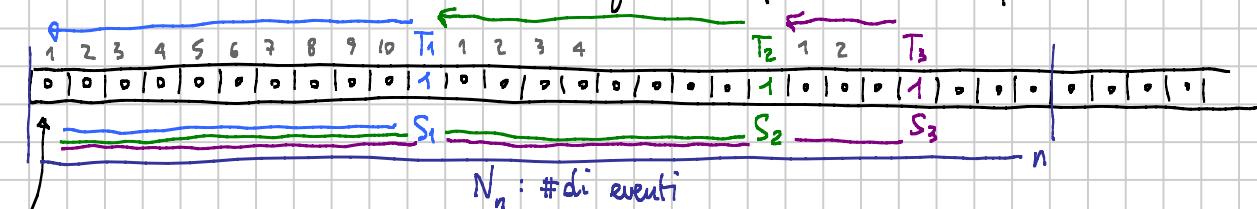
→ Non è riproducibile: se si sommano vari iid di questo tipo...



→ Momenti: $E(X) = \frac{n+1}{2}$ $\text{Var}(X) = \frac{n^2-1}{12}$

• GEOMETRICA, BINOMIALE NEGATIVA, GAMMA-POISSON

→ Processo di Bernoulli: analogo a tempi discreti del processo di Poisson



qui casella 1 con prob p e 0 con prob $1-p$, indip.

$N_n \sim \text{bin}(n, p)$

$T_i \sim \text{geom}(p)$

$S_m \sim \text{negbin}(m, p)$

* Normalmente con negbin(m, p) si intende il numero di prove prima del successo m -esimo (lui compreso)

$$P(\text{negbin}(m, p) = k) = \binom{k-1}{m-1} p^m (1-p)^{k-m} \quad k = m, m+1, m+2, \dots$$

* Quella di `scipy.stats.nbinom` e della Gamma-Poisson intende il numero di fallimenti prima del successo n-esimo

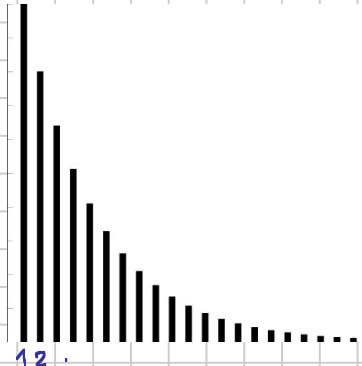
• GEOMETRICA

$$X \sim \text{geom}(p)$$

$$p \in (0, 1]$$

$$p=0 \Rightarrow X = +\infty$$

$$P(X=k) = p(1-p)^{k-1} \quad k=1, 2, 3, \dots$$



→ si trova quando c'è un processo di Bernoulli

$$\rightarrow X \sim \text{geom}(p) \quad E(X) = \frac{1}{p} \quad \text{Var}(X) = \frac{1-p}{p^2}$$

→ è la versione discreta dell'esponenziale

•

BINOMIALE NEGATIVA

$$X \sim \text{negbin}(r, p)$$

$$n \text{ intero positivo } p \in (0, 1]$$

$$P(X=k) = \binom{k-1}{r-1} p^r (1-p)^{k-r} \quad k \geq r$$

→ conta il numero di prove prima dell' r -esimo successo in un processo di Bernoulli di parametro p

→ $r=1$: geometrica

→ è riproducibile : X_1, \dots, X_m indip. $X_i \sim \text{negbin}(r_i; p)$
allora $X_1 + \dots + X_m \sim \text{negbin}(r_1 + \dots + r_m; p)$

$$\rightarrow X \sim \text{negbin}(r, p) \quad E(X) = \frac{r}{p} \quad \text{Var}(X) = r \cdot \frac{1-p}{p^2}$$

→ esiste un'altra definizione, dove $\tilde{X} \sim \text{negbin}(r, p)$ nel senso invece che \tilde{X} è il numero di insuccessi e non di prove : $\tilde{X} := X - r$

$$P(\tilde{X}=k) = P(X=k+r) = \binom{r-1+k}{r-1} p^r (1-p)^k = \frac{\Gamma(r+k)}{k! \Gamma(r)} (1-p)^k p^r \quad k \geq 0 \quad (\text{vedi Gamma-Poisson})$$

GAMMA - Poisson

 $X \sim gp(r, p) \circ X \sim gp(\nu, a)$ due diverse parametrizzazioni

Distribuzione a due parametri sugli interi non negativi

https://en.wikipedia.org/wiki/Negative_binomial_distribution#Gamma%E2%80%93Poisson_mixture

→ si ottiene generalmente prima una $T \sim \text{gamma} (\alpha=r, \lambda=\frac{p}{1-p})$
e successivamente una $X \sim \text{Pois}(T)$

$$\rightarrow E(T) = \frac{\alpha}{\lambda} = r \frac{1-p}{p}$$

$$\Rightarrow E(X) = E(T) = r \frac{1-p}{p}$$

$$\begin{aligned} &= \nu \\ &= \nu + a\nu^2 \end{aligned}$$

$$\begin{cases} \nu = r \frac{1-p}{p} \\ a = \frac{1}{r} \end{cases} \quad \begin{cases} p = \frac{1}{1+\nu} \\ r = \frac{1}{\nu} \end{cases}$$

* NB: $\text{Var}(X) \geq E(X)$

$$P(X=k) = \frac{\Gamma(r+k)}{k! \Gamma(r)} (1-p)^k p^r \quad k=0,1,\dots$$

per la funzione di massa di prob.

→ Generalizza la Poisson (con varianza maggiore)
per $a \rightarrow 0$ si ottiene la legge di Pois(ν)

The probability mass function of the number of failures for `nbinom` is:

$$f(k) = \binom{k+n-1}{n-1} p^n (1-p)^k \quad k=0,1,\dots \quad \frac{(k+n-1)!}{(n-1)! \cdot k!} = \frac{\Gamma(k+n)}{\Gamma(n) \cdot k!}$$

for $k \geq 0, 0 < p \leq 1$ [9] `nbinom.pmf(10, n=[3, 3.5, 4], p=0.2)`

array([0.05669357, 0.05448781, 0.04913443])

$$n! = \Gamma(n+1)$$

$$\frac{n!(n+1)}{(n+1)!} = \frac{(n+1)!}{(n+1)!} = 1$$

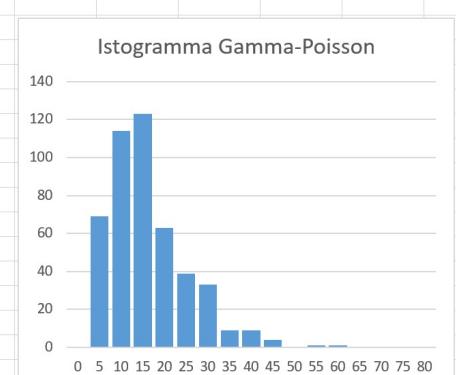
scipy la gestisce bene

→ Generalizza la binomiale negativa (con ν non intero)

3.5 alfa=r=n
4 beta=1/lambda=(1-p)/p
14 media
0.2 p
70 varianza
0.28571 a

gamma-poisson

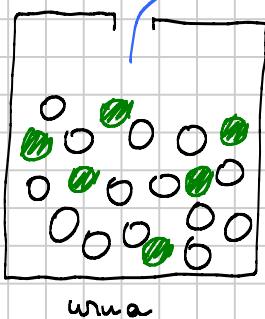
gamma	gamma-poisson	21	0	0
11.3367	8	19	35	9
23.7364	25	20	40	9
3.06647	6	15	45	4
19.4184	14	9	50	0
17.9181	14	5	55	1
7.89043	7	21	60	1
15.8867	6	23	65	0
14.195	10	17	70	0
20.5857	16	17	75	0



IPERGEOMETRICA

$$X \sim \text{hypg}(n, a, m)$$

$$m \geq 1 \quad 0 \leq a \leq m \\ 1 \leq n \leq m$$



$$(n, a, m) \quad E(x) = \frac{na}{m}$$

$\text{Var}(x)$ poco meno di quella della bin corrispondenti

m : totale di palline nell'urna

a : palline colorate

n : quante ne estraggo senza riwersa

X : quante colorate ho estratto

$$P(X = k) = \frac{\binom{a}{k} \binom{m-a}{n-k}}{\binom{m}{n}}$$

$$\max(0, n - (m-a)) \leq k \leq \min(a, n)$$

→ nel caso con riwersa $X \sim \text{bin}(n, \frac{a}{m})$ invece (diverso!)

→ se m è molto grande, anche rispetto a n , $X \sim \text{bin}(n, \frac{a}{m})$

VALORE ATTESO

aka: media di una v.a.

$$E(x) = \int_{\mathbb{R}} x f_x(x) dx \quad \rightarrow$$

→ è il baricentro

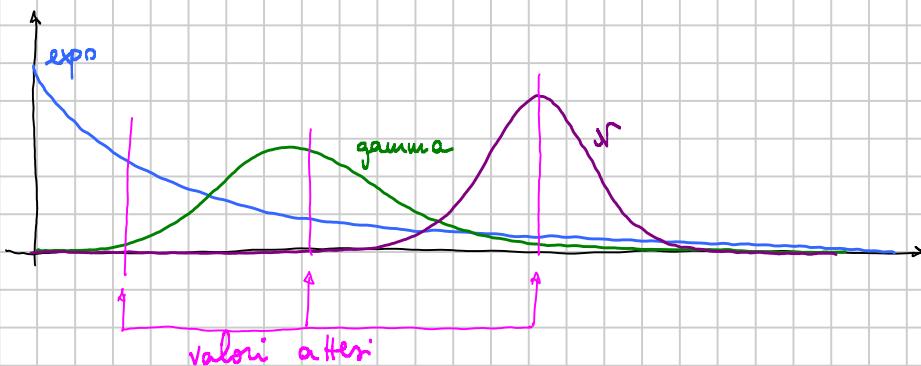
valore atteso di una funzione di x

$$E(g(x)) = \int_{\mathbb{R}} g(x) f_x(x) dx$$

$$E(g(x)) = \sum_k g(k) \varphi_x(k)$$

* il valore atteso è una delle misure di centralità di una distribuzione

↳ l'altra è la mediana



* Il valore atteso è lineare :

$$\forall \alpha, \beta_1, \dots, \beta_n \in \mathbb{R}$$



$$E(\alpha + \beta x) = \alpha + \beta E(x)$$

$$E(x_1 x_1 + \alpha_2 x_2 + \dots + \alpha_n x_n) = \alpha_1 E(x_1) + \alpha_2 E(x_2) + \dots + \alpha_n E(x_n)$$

- Pro e contro tra media e mediana $E(x) \quad M(x)$
 - la mediana non risente di trasformazioni monotone
 $M(\log(x)) = \log(M(x))$ $M(g(x)) = g(M(x))$ g monotona
 - la media in generale sì: non risente solo di cambiamenti di scala
 $E(32 + \frac{9}{4}T) = 32 + \frac{9}{4}E(T)$ (è lineare!)
 - la mediana minimizza gli scarti / la distanza media
 $M(x) = \underset{t \in \mathbb{R}}{\operatorname{argmin}} E(|x - t|)$
 $M_x = \underset{t \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^n |X_i - t|$ (mediana campionaria)
 - la media minimizza gli scarti / la distanza media
 $E(x) = \underset{t \in \mathbb{R}}{\operatorname{argmin}} E((x - t)^2)$
 $\bar{x} = \underset{t \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^n (X_i - t)^2$ (media campionaria)
- * Conviene sempre portarsi a leggi simmetriche (come da logaritmale a \mathcal{N}) perché media e mediana coincidono, e possiamo usare la teoria sulla prima per dire anche cose sulla seconda.

HW: cercare la definizione di mediana campionaria

• VALORE ATTESO

aka : media di una v.a,

$$E(x) = \int_{\mathbb{R}} x f_x(x) dx \rightarrow$$

→ è il baricentro

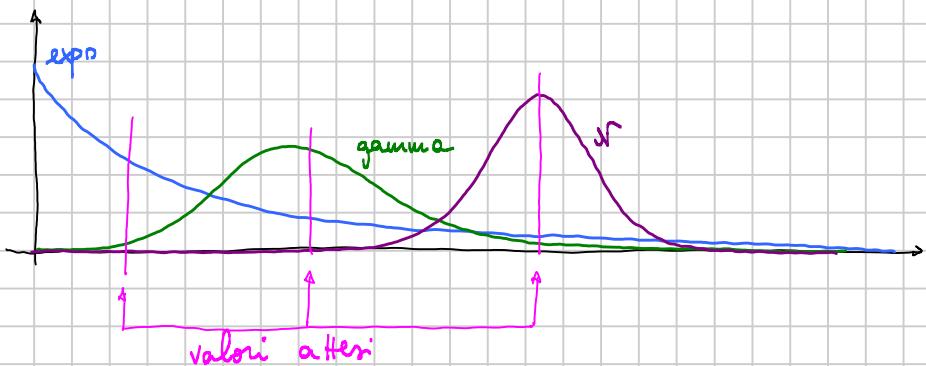
valore atteso di una funzione di x

$$E(g(x)) = \int_{\mathbb{R}} g(x) f_x(x) dx$$

$$E(g(x)) = \sum_k g(k) \varphi_x(k)$$

* il valore atteso è una delle misure di **centralità** di una distribuzione

↳ l'altra è la **mediana**



* Il valore atteso è lineare : $X, E(x) \rightsquigarrow \alpha X + \beta, \alpha E(x) + \beta$ 🏁

(e il baricentro : segue la distribuzione)

• Momenti successivi

i. Varianza $\text{Var}(x) := E[(x - E(x))^2] = E(x^2) - E(x)^2$

$$E(x^2) = \int_{\mathbb{R}} x^2 f_x(x) dx$$

→ momento secondo centrato → è il momento di inerzia

→ $\sqrt{\text{Var}(x)}$ è la deviazione standard

* La deviazione standard è una delle misure di **dispersione** di una distribuzione

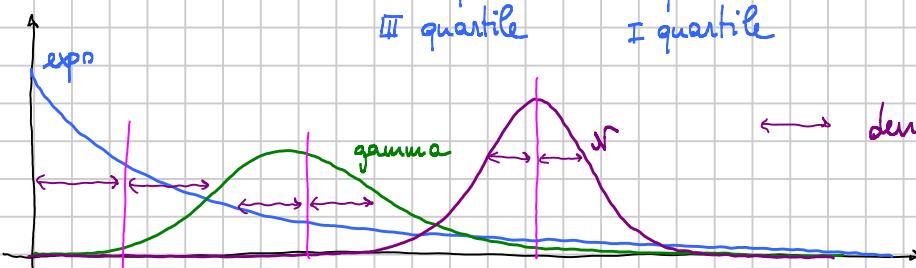
↳ un'altra è il range interquartile

$$q_3 - q_1 := F^{-1}(75\%) - F^{-1}(25\%)$$

III quartile

I quartile

↔ deviazioni standard



③ MOMENTI E TRASFORMAZIONI LINEARI : SKEWNESS & KURTOSI

a) Momento primo : valore atteso $E(x)$

- risente dello **shift** e della **scala** : $E(a+bx) = a+bE(x)$

- rappresenta il **centro** della distribuzione

★ Tutti i momenti **centrati** sono fatti in modo da **non risentire dello shift**

b) Momento secondo centrato : varianza $Var(x)$

- non risente dello **shift**, ma solo della **scala** : $Var(a+bx) = b^2 Var(x)$

- rappresenta la **larghezza** della distribuzione

c) Momento terzo centrato e standardizzato : skewness $sk(x)$

$$sk(x) = \frac{E[(x-E(x))^3]}{Var(x)^{3/2}} \in \mathbb{R}$$

- non risente di **shift** né **scala** $sk(a+bx) = sk(x)$ (HW)

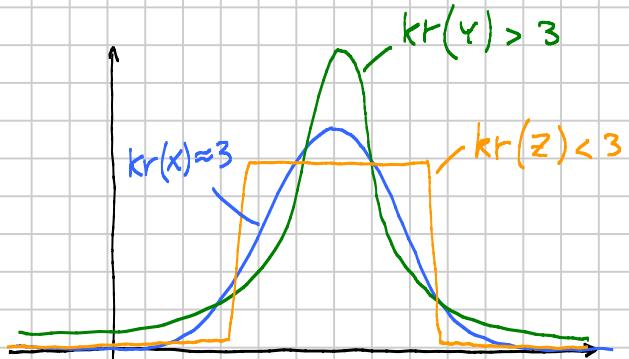
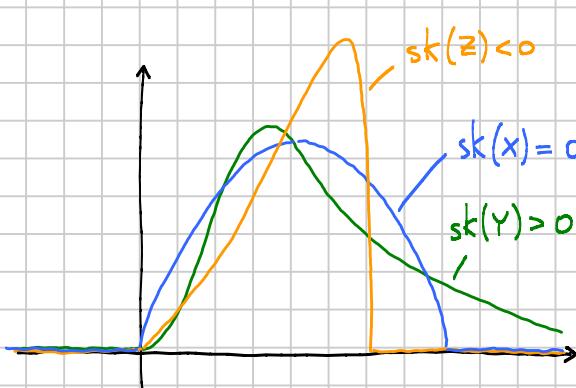
- dipende dalla **forma : assimetria** ($\neq 0$ per distrib. simmetriche)

d) Momento quarto centrato e standardizzato : kurtosi $kr(x)$

$$kr(x) = \frac{E[(x-E(x))^4]}{Var(x)^2} \geq 0$$

- non risente di **shift** né **scala** $kr(a+bx) = kr(x)$ (HW)

- dipende dalla **forma : modo in cui è concentrata la distribuzione**



↳ misura di schiacciamento

↳ con questa definizione se $X \sim N$ $kr(x) = 3$

quindi per alcuni autori $k_f(x) = kr(x) - 3$

* Nel caso di $sk(x)$ e $kr(x)$ prima di calcolare i momenti terzo e quarto ho standardizzato la v.a. ovvero ho applicato quell'unica trasformazione lineare che punta la media a 0 e la dev. std a 1

$$E\left(\frac{X-\mu}{\sigma}\right) = \frac{E(X)-\mu}{\sigma} = \frac{\mu-\mu}{\sigma} = 0$$

$$\text{Var}\left(\frac{X-\mu}{\sigma}\right) = \text{Var}\left(\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma^2} \text{Var}(X) = \frac{\sigma^2}{\sigma^2} = 1$$

$\uparrow \alpha \quad \uparrow \beta$

→ secondo alcuni autori, questa operazione si chiama normalizzazione

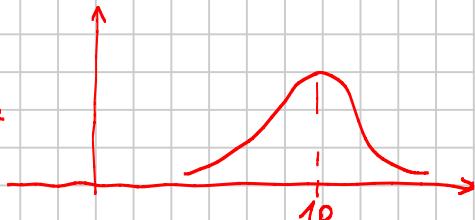
↳ a causa di ciò skewness e kurtosis sono invarianti per trasformazioni lineari: sono quindi detti parametri di forma / shape

$$sk(a+bX) = sk(X) \quad kr(a+bX) = kr(X)$$

HW: Consideriamo la legge $t(4)+10$

Determinare media μ e varianza σ^2

(posizione e lunghezza)



Sorapporre sul grafico delle pdf le pdf di:

uniforme, Gaussiana, gamma di media μ e varianza σ^2

Cercare sk e Kr di tutte e quattro

Determinazione approssimata di valori attesi e probabilità

→ vedremo come si risolve con simulazione Monte Carlo e integrale numerico o esaurizione di casi

Esempi

- a. Successione infinita di $\text{unif}(0,1)$ iid

$$U_1, U_2, \dots, U_n, \dots \quad U_i \sim \text{unif}(0,1)$$

consideriamo le somme parziali o cumulate

$$S_1 = U_1$$

$$S_2 = U_1 + U_2$$

$$S_n = \sum_{i=1}^n U_i$$

$$S_3 = U_1 + U_2 + U_3$$

avremo $0 \leq S_1 \leq S_2 \leq \dots$ $\lim_{n \rightarrow \infty} S_n = \infty$ HW: perché? (hint: LGN)

mi domando quanto "tempo" ci vuole per superare una soglia fissata:

$$\alpha > 0 \text{ soglia} \quad T := \inf \{ n : S_n \geq \alpha \}$$

T è una v.a. discreta che assume valori interi

la distribuzione di T è complicata e dipende da α

→ come stimo $E(T)$? come stimo $P(T \leq c)$?

* Se $\alpha = 1$ mi dimostra che $E(T) = e$

b. $Z \sim \mathcal{N}(\mu, \sigma^2)$ $E(\sqrt{Z}) = ?$ X non ha senso $E(\sqrt{Z} \cdot \mathbb{1}_{Z \geq 0}) = E\left(\begin{cases} \sqrt{Z} & Z \geq 0 \\ 0 & Z < 0 \end{cases}\right) = ?$

$$\int_0^\infty \sqrt{x} \cdot f_Z(x) dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^\infty \sqrt{x} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = ?$$

c. $X \sim \text{negbin}(r, p)$ $r = 2.5$ $p = 0.1$ Gamma-Poisson

Voglio calcolare l'entropia:

$$H(x) := E[-\log(\varphi_x(x))] = -\sum_{k \geq 0} \varphi_x(k) \log \varphi_x(k)$$

d. battaglia nel Risiko (3 armate contro 3 armate)

tre dadi rossi per l'attaccante : 5, 2, 6 $2 \leq 5 \leq 6$
 tre dadi blu per il difensore : 1, 4, 6 $1 \leq 4 \leq 6$ $\rightarrow 2-1$ per attaccante

Esorti possibili : $\begin{matrix} 3-0 \\ 2-1 \\ 1-2 \\ 0-3 \end{matrix}$ } quali sono le
 } probabilità di questi esiti?

• Soluzione stocastica : simulazione Monte Carlo

genero molte volte le varie coinvolte ($N \approx 1k$ o più)

\hookrightarrow stimo i valori attesi con le medie campionarie

X con una certa legge $\rightarrow E(x) = ?$

X_1, X_2, \dots, X_N generate con S.M.C. $\bar{X} := \frac{1}{N} \sum_{i=1}^N X_i \approx E(x)$ per la L.G.N

$\text{Var}(\bar{X}) = \frac{1}{N} \text{Var}(x)$ la varianza tende a zero per $N \rightarrow \infty$

$$\begin{array}{ll} \text{a. } U_{1,1} U_{1,2} U_{1,3} \dots & U_{1,K} \left| \begin{array}{c} S_{1,1} S_{1,2} \dots S_{1,K} \\ T_1 \\ \vdots \\ T_N \end{array} \right| \leq C \\ U_{2,1} U_{2,2} U_{2,3} \dots & U_{2,K} \left| \begin{array}{c} S_{2,1} \dots \\ T_2 \\ \vdots \\ T_N \end{array} \right| \\ \dots & \vdots \\ U_{N,1} U_{N,2} \dots & U_{N,K} \left| \begin{array}{c} S_{N,1} \\ T_N \end{array} \right| \end{array}$$

campione finale

\bar{T}

$$\begin{cases} 1 \\ 0 \\ \vdots \\ 0 \end{cases}$$

media

$$f_C = \frac{1}{N} \sum_{i=1}^N I_{T_i \leq C}$$

$$E(T) \approx \bar{T} = \frac{1}{N} \sum_{i=1}^N T_i$$

* Si può anche fare l'intervalle di confidenza

— Recall —

$$x_1, \dots, x_N \sim \mathcal{N}(\mu, \sigma^2) \quad \mu = E(x) \quad \bar{x} = \mu$$

$$\mu \in \bar{x} \pm q \frac{s_x}{\sqrt{N}} \quad q = F_{t(N-1)}^{-1} (1 - \frac{\alpha}{2}) \text{ con livello di conf } 1 - \alpha$$

$$\rightarrow \text{per } N \gg 1 \quad t(N-1) \sim \mathcal{N}(0,1) \quad q \approx \Phi^{-1} (1 - \frac{\alpha}{2})$$

$$\rightarrow \text{per } \alpha \approx 5\% \quad q \approx 2$$

$$E(x) \in \bar{x} \pm q \frac{s_x}{\sqrt{N}} \text{ al livello di conf } 1 - \alpha$$

$$E(\bar{T}) \in \bar{T} = 2 \cdot \frac{S_T}{\sqrt{N}} \quad \text{dove} \quad S_T := \sqrt{\frac{1}{N-1} \sum_{i=1}^N (T_i - \bar{T})^2}$$

→ Se l'intervallo sembra troppo largo, si può aumentare N .

* anche probabilità

$$P(T \leq c) = ? \quad \approx \frac{\#\{i : T_i \leq c\}}{N} =: f_c$$

↳ in pratica abbiamo stimato la media di

$$\mathbb{1}_{T \leq c} := \begin{cases} 1 & T \leq c \\ 0 & \text{altrimenti} \end{cases}$$

$$\mathbb{1}_{T \leq c}$$

$$* E(\mathbb{1}_A) = P(A)$$

HW: check (hint: $\mathbb{1}_A$ è di Bernoulli)

↳ si può fare l'intervallo di confidenza:

$$p \in \hat{p} \pm q \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$P(T \leq c) \approx f_c \pm q \sqrt{\frac{f_c(1-f_c)}{N}}$$

q come sopra

b. $Z_1, Z_2, \dots, Z_N \sim \mathcal{N}(\mu, \sigma^2)$ generate dalla SMC

$$\text{calcolo } X_1 = \sum Z_i \mathbb{1}_{Z_i \geq 0} \quad X_2 = \sum Z_i \mathbb{1}_{Z_i \geq 0} \dots \quad X_N = \sum Z_i \mathbb{1}_{Z_i \geq 0} \quad E(\bar{Z} \mathbb{1}_{Z_i \geq 0}) \approx \bar{X} \pm q \frac{S_x}{\sqrt{N}}$$

c. $H(x) = -E(\log \varphi_x(x)) =: E(Y) \approx \bar{Y} \pm q \frac{S_Y}{\sqrt{N}}$

$$Y := -\log \varphi_x(x)$$

$$\text{SMC: } x_i \sim nb(r, p) \quad i=1, 2, \dots, N \rightarrow Y_i = -\log \varphi_x(x_i)$$

						HW: provare, poi si ride
$A_{1,1}$	$A_{1,2}$	$A_{1,3}$	$D_{1,1}$	$D_{1,2}$	$D_{1,3}$	$A_{1,(1)}$ $A_{1,(2)}$ $A_{1,(3)}$ $D_{1,(1)}$ $D_{1,(2)}$ $D_{1,(3)}$ $X_1 \in \{0, 1, 2, 3\}$
$A_{2,1}$...		$D_{2,3}$			
:			:			
$A_{N,1}$			$D_{N,3}$		X_N

$$P(X=3) \approx f_3 \pm q \sqrt{\frac{f_3(1-f_3)}{N}} \quad f_3 := \frac{\#\{i : x_i = 3\}}{N}$$

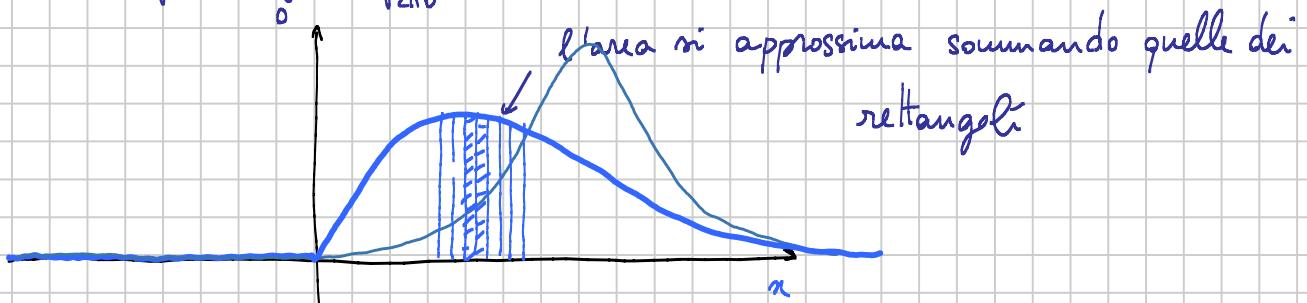
* In tutti questi casi è utile e opportuno anche controllare qualitativamente la distribuzione del campione.

• Soluzione numerica : integrale numerico o esaurizione dei casi

a. IMPOSSIBILE

b. Integro

$$\int_0^\infty \sqrt{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$



c. Sommo $\sum_{k=0}^{\infty} -\log \varphi(k) \cdot \varphi(k)$

→ invece di sommare fino a $k \rightarrow \infty$ mi fermerò quando gli addendi diventano piccoli

d.

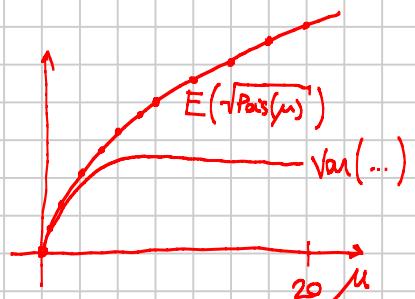
1	1	1	1	1	1	x
1	1	1	1	1	2	0
...	...					0
6	6	6	6	6	6	:

infatti tutti i casi possibili sono $6^6 = 46656$

HW.1 Stimare con entrambi i metodi

$E(\sqrt{X})$, $\text{Var}(\sqrt{X})$ dare $X \sim \text{Pois}(\mu)$

Chiedo il grafico in funzione di μ



HW.2 Sia Y una v.a. positiva approssimativamente

Gaussiană $Y \sim \mathcal{N}(\mu, \sigma^2)$

$\mu > \sigma$



Stimare $E(\sqrt{Y})$

in funzione di μ e σ

(Possibilmente capire come ridurre da due a un parametro senza perdita di generalità)

INTRODUZIONE AL MACHINE LEARNING

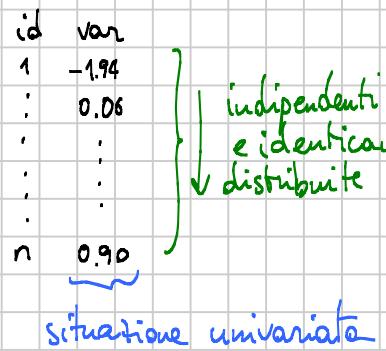
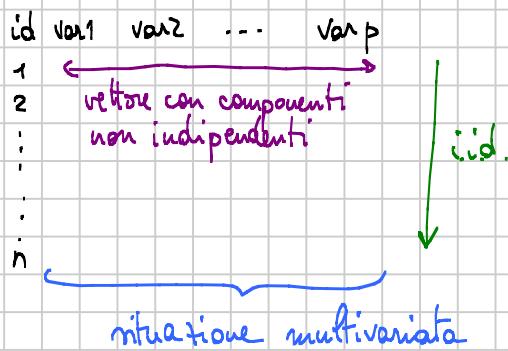
ord 7

04/03/2025

VETTORI ALEATORI

→ situazione standard da M.L. o data science

il dataset di partenza è una tabella



- * Serve la teoria dei vettori aleatori, ma leggendo: lavoriamo quasi solo con i primi due momenti

$X = (X_1, X_2, \dots, X_m)$ un vettore con componenti casuali

(ad esempio: 6 dadi del Rischio ($m=6$, indipendenti))

- o gli stessi 6 dadi ordinati $X_{(1)} \leq X_{(2)} \leq X_{(3)} \dots$ (non indip.) .)

1) $\boxed{\mu := E(X)}$ è un vettore deterministico



$\boxed{\mu_i := E(X_i)}$

$\mu \in \mathbb{R}^m \quad \mu = (\mu_1, \mu_2, \dots, \mu_m)$

media componente per componente

2) $\Sigma := C(X)$ è una matrice $\Sigma \in M_{m,m}$

$$\Sigma = \begin{pmatrix} \sum_{i=1}^n \sum_{j=1}^{i+1} \dots \sum_{j=m}^m \\ \vdots \\ \sum_{i=1}^n \sum_{j=1}^{i+1} \dots \sum_{j=m}^m \end{pmatrix}$$

$\boxed{\Sigma_{ij} := Cov(X_i; X_j)}$ matrice di covarianza

HW: $Y = (Y_1, Y_2, \dots, Y_6) = (X_{(1)}, X_{(2)}, \dots, X_{(6)})$ \times dadi indipendenti (sei uguali)
stimare $E(Y)$ (un vettore $(\mu_1, \mu_2, \dots, \mu_6)$) e $C(Y)$

Recall : $Cov(X; Y) := E[(X - E(X)) \cdot (Y - E(Y))]$

$$Cov(X; X) = Var(X)$$

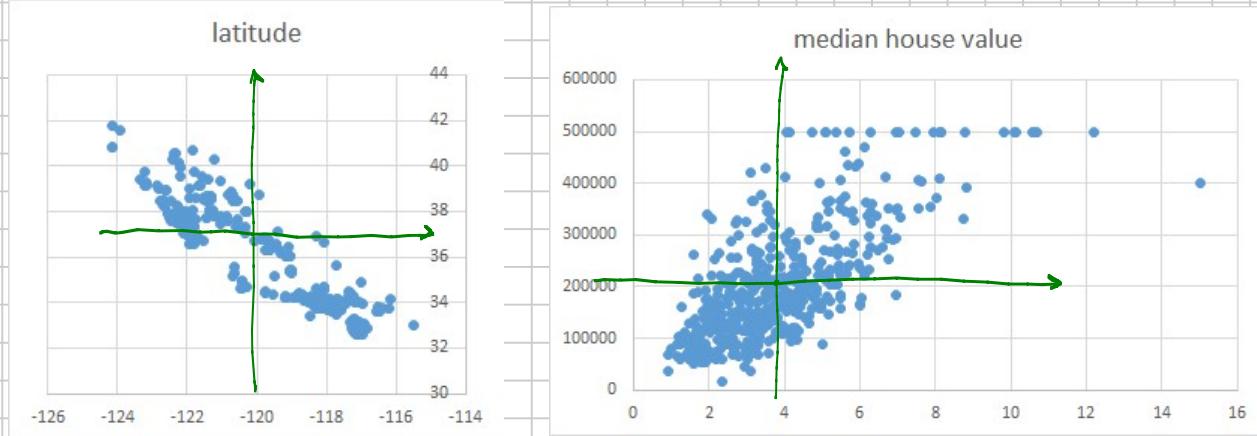
$$X, Y \text{ indip} \Rightarrow Cov(X, Y) = 0$$

$$Cov(X; Y) = Cov(Y; X)$$

$$Cov(\text{cost}; X) = Cov(\text{cost}; \text{cost}) = 0$$

* Quindi $C(x)$ è una matrice simmetrica e sulla diagonale ci sono le varianze delle x_i (sempre positive)

$$C(x) = \begin{pmatrix} \text{Var}(x_1) & \text{Cov}(x_1, x_2) & \dots & \text{Cov}(x_1, x_m) \\ \text{Cov}(x_2, x_1) & \text{Var}(x_2) & & \vdots \\ \ddots & \ddots & \ddots & \vdots \\ \ddots & \ddots & \ddots & \text{Var}(x_m) \end{pmatrix}$$



$$\text{Cov}(x, Y) < 0$$

$$\text{Cov}(x, Y) > 0$$

🚩 Se le componenti x_1, \dots, x_n sono indipendenti, $C(x)$ è diagonale

Recall: la covarianza è bilineare

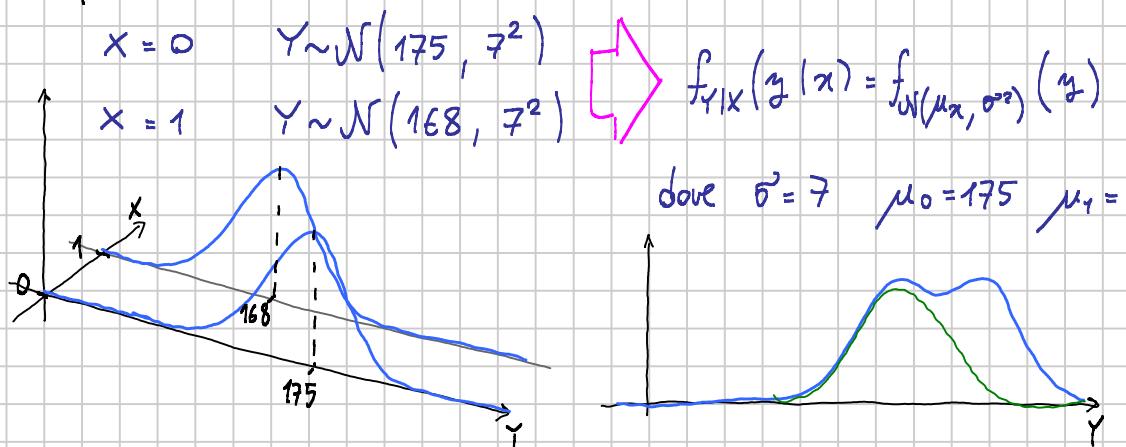
$$\text{Cov}\left(\sum_{i=1}^m a_i x_i ; \sum_{j=1}^n b_j Y_j\right) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \text{Cov}(x_i ; Y_j)$$

• Legge congiunta nel caso misto: un esempio per capire

due componenti: X : sesso Y : statura (uomini adulti)

$\rightarrow X$ discreta $\in \{0, 1\}$ 0: maschio 1: femmina $P(X=1) = \frac{1}{2}$

$\rightarrow Y$ continua, Gaussiana sia per i maschi, sia per le femmine con parametri diversi



■ Vettori alatori e ML

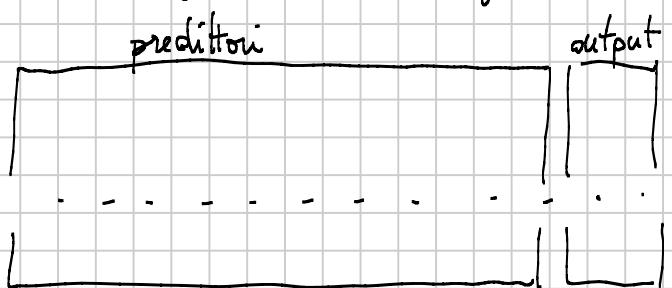
1) Supervised Learning : prevedo delle variabili di output a partire da quelle di input

a) prevedere risorse necessarie (tempo, soldi, energia)
per diverse commesse / progetti

b) prevedere il peso corporeo date altre misure

c) prevedere categoria di un'immagine

di solito una variabile singola



d) sesso → statura

dopo aver imparato la relazione fra X e Y

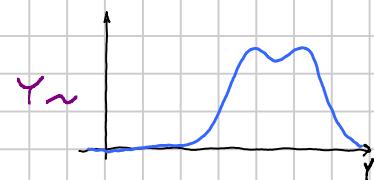
$$X = 0 \rightarrow Y \approx 175 \quad Y \sim N(175; 7^2)$$

$$X = 1 \rightarrow Y \approx 168 \quad Y \sim N(168; 7^2)$$

* Se non c'è dipendenza tra le variabili la predizione è la distribuzione di Y

Z : 0/1 occhi scuri/chiaro

$$\begin{aligned} Z = 0 &\rightarrow Y \\ Z = 1 &\rightarrow Y \end{aligned} \quad \approx 171.5$$



2) Unsupervised Learning : cerco di capire com'è la distribuzione multivariata del dataset

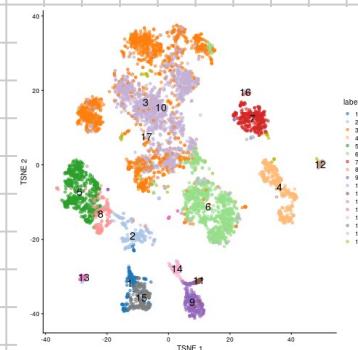
e) clusterizzazione cellule : righe sono cellule (5k)

colonne sono geni (15k)

→ trovo le relazioni fra le variabili

→ scopro come si raggruppano le cellule

→ cerco di identificarle

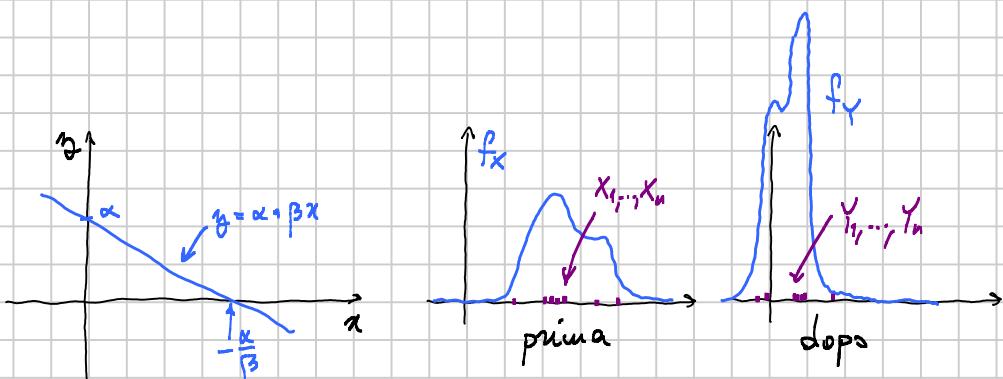


■ TRASFORMAZIONI LINEARI DI VETTORI ALEATORI

Funzione lineareda $\mathbb{R} \rightarrow \mathbb{R}$: retta

$$y = \alpha + \beta x$$

2 parametri



→ trasformazione della legge teorica e del campione empirico

ci sarà un po' di imprecisione / vaghezza nel seguito, su questo

legge teoria - valore atteso - covarianza - mediana ...

$$f_x$$

$$E(x)$$

$$C(x)$$

$$M_x$$

campione - media campion. - covarianza c. - mediana c.

$$x_1, \dots, x_n$$

$$\bar{x}$$

$$\sum_{\bar{x}}$$

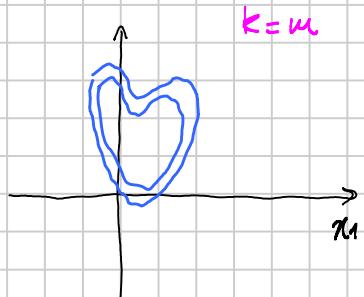
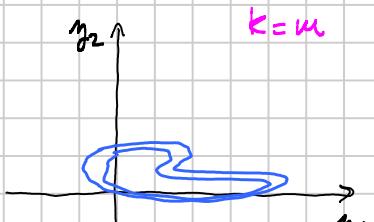
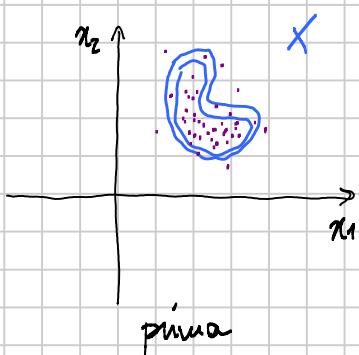
?

- Caso vettoriale : x vettore aleatorio $y = g(x)$ g lineare

da $\mathbb{R}^m \rightarrow \mathbb{R}^k$: $\vec{y} = \vec{\alpha} + B\vec{x}$ $\vec{x} \in \mathbb{R}^m$ $\vec{y}, \vec{\alpha} \in \mathbb{R}^k$, $B \in M_{k,m}$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} + \begin{pmatrix} b_{11} & \dots & b_{1m} \\ b_{21} & \dots & b_{2m} \\ \vdots & \ddots & \vdots \\ b_{k1} & \dots & b_{km} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

$$\begin{cases} y_1 = \alpha_1 + b_{11}x_1 + b_{12}x_2 + \dots + b_{1m}x_m \\ \dots \\ y_k = \alpha_k + b_{k1}x_1 + \dots + b_{km}x_m \end{cases}$$

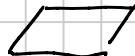


dopo : ogni componente
trasformata da sola

B diagonale

$$\begin{cases} y_1 = \alpha_1 + b_{11}x_1 \\ y_2 = \alpha_2 + b_{22}x_2 \end{cases}$$

dopo : generali
rotazioni più
deformazioni



• Come si trasformano media e covarianza

$$X \text{ vettore aleatorio di dim } m \quad \underline{\mu_X = E(X) \in \mathbb{R}^m} \quad \underline{\Sigma_X = C(X) \in M_{m,m}}$$

trasformazione lineare $\boxed{Y = \alpha + BX}$ Y vettore aleatorio a valori in \mathbb{R}^k

$\alpha \in \mathbb{R}^k$ $B \in M_{k,m}$ assegnati

$$\mu_Y = E(Y) = E(\alpha + BX) = \alpha + B\mu_X \quad \text{linearità}$$

$$Y_i = \alpha_i + [BX]_i = \alpha_i + \sum_{j=1}^m B_{ij} X_j$$

$\underline{[\mu_Y]_i}$

$$[\mu_Y]_i = [E(Y)]_i = E(Y_i) = E\left(\alpha_i + \sum_j B_{ij} X_j\right) = \alpha_i + \sum_j B_{ij} E(X_j) = \alpha_i + [B\mu_X]_i = [\alpha + B\mu_X]_i$$



$$\boxed{\mu_Y = \alpha + B\mu_X}$$

$$\boxed{\bar{Y} = \alpha + B\bar{X}}$$

$$\Sigma_Y = C(Y) = C(\alpha + BX) = B C(X) B^T$$

$$\text{Var}(\alpha + \beta X) \propto \beta^2 \text{Var}(X)$$

$$[\Sigma_Y]_{ij} = [C(Y)]_{ij} = \text{Cov}(Y_i; Y_j) = \text{Cov}\left(\alpha_i + \sum_{k=1}^m B_{ik} X_k; \alpha_j + \sum_{h=1}^m B_{jh} X_h\right)$$

(bilinearità)

$$= \sum_k \sum_h B_{ik} B_{jh} \text{Cov}(X_k; X_h) = \sum_{k,h} B_{ik} [\Sigma_X]_{kh} [B^T]_{hj} = [B \Sigma_X B^T]_{ij}$$



$$\boxed{\Sigma_Y = B \Sigma_X B^T}$$

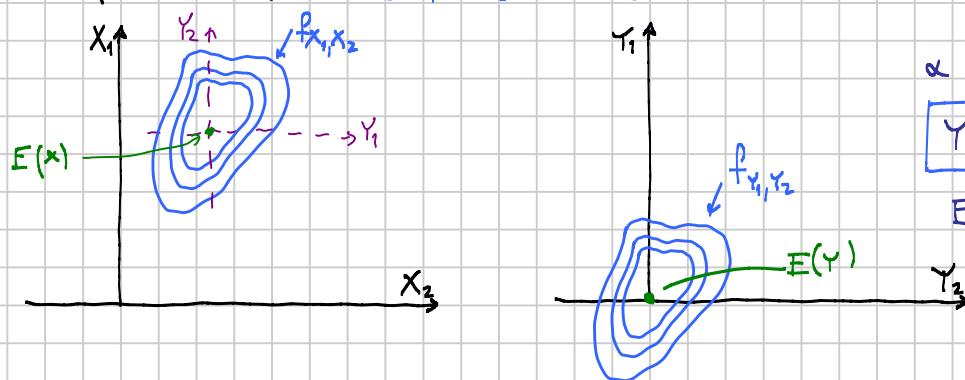
* Inoltre se $m=k$, B invertibile e X ha pdf $f_X: \mathbb{R}^m \rightarrow \mathbb{R}_+$

Allora Y ha una pdf in dim m $f_Y: \mathbb{R}^m \rightarrow \mathbb{R}$

$$\boxed{f_Y(y) = |\det(B)|^{-1} f_X(B^{-1}y)}$$

TRASFORMAZIONI LINEARI FREQUENTI

- trasformazione che centra il vettore



$$\alpha := -\mu_X \quad B = I \in M_{m,m}$$

$$Y = X - \mu_X \quad E(Y) = 0 \quad C(Y) = \sum_X$$

→ per ciascuna componente, si sottrae la media a livello di campione?

$$X_{ij} - \bar{x}_j \rightarrow Y_{ij}$$

id	var1	var2	...	var m
1	x_{11}	x_{12}		x_{1m}
2	:			:
:	:			:
:				
n	x_{n1}		x_{nm}

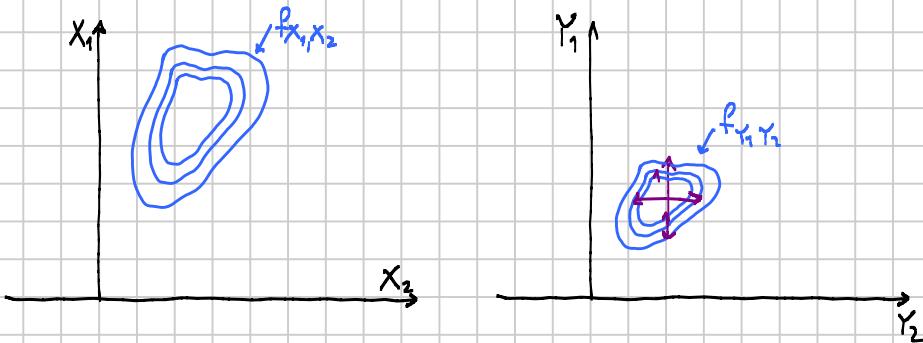
X_{ij} : dato var j id i

$$\begin{cases} \bar{x}_1 & \bar{x}_2 \\ \bar{x}_{n1} & \bar{x}_{n2} \end{cases} \quad \begin{cases} \bar{x}_m \\ \bar{x}_{nm} \end{cases}$$

$$\bar{x}_j = \bar{x}_{*j} = \frac{1}{n} \sum_{i=1}^n X_{ij} \quad \text{media campionaria comp } j$$

due notazioni standard

- trasformazione che standardizza le variante (var → 1)



$$Y = \alpha + BX$$

$$B = \begin{pmatrix} \text{std}(x_1)^{-1} & & & & 0 \\ & \text{std}(x_2)^{-1} & & & \\ & & \ddots & & \\ 0 & & & \ddots & \text{std}(x_m)^{-1} \end{pmatrix}$$

$$B = \text{diag}(\text{std}(x)^{-1})$$

$$X_i / \text{std}(x_i) \rightarrow Y_i$$

$$\text{std}(x_i) = \sqrt{\text{Var}(x_i)} \quad \text{deviazione standard}$$

$$X_{ij} / S_{x_j} \rightarrow Y_i$$

$$S_{x_j} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_{ij} - \bar{x}_j)^2} \quad \text{dev. std campionaria comp } j.$$

→ Cosa succede a μ_Y e Σ_Y ?

$$\mu_Y = \alpha + B\mu_X = \text{diag}(\text{std}(x)^{-1})\mu_X \quad E(Y_i) = E(X_i)/\text{std}(X_i)$$

$$\Sigma_Y = B\Sigma_X B^T = \text{diag}(\text{std}(x)^{-1})\Sigma_X \text{diag}(\text{std}(x)^{-1})$$

$$C(Y)_{ij} := C(X)_{ij} / (\text{std}(X_i) \cdot \text{std}(X_j)) = \frac{\text{Cov}(X_i; X_j)}{\sqrt{\text{Var}(X_i)\text{Var}(X_j)}} =: \rho(X_i; X_j)$$

coeff di corr lineare

→ La matrice di covarianza di Y è la matrice di correlazione lineare di X

$$C(Y) = \begin{pmatrix} 1 & \rho(X_1, X_2) & \dots & \rho(X_1, X_m) \\ \rho(X_2, X_1) & 1 & & \vdots \\ \vdots & & \ddots & \rho(X_m, X_m) \\ \vdots & \dots & \dots & 1 \end{pmatrix} =: \rho(X) \quad \text{Hw: } \rho(Y) = \rho(X)$$

- def Il coefficiente di correlazione lineare $\rho(X, Y)$ di due waa X e Y è
(se esistono $\text{Var}(X), \text{Var}(Y) < \infty$) :

Ross pblm 4.51



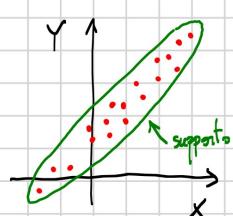
$$\rho(X, Y) := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

$\in [-1; 1]$

$\rho(X, Y)$ misura alla covarianza nella scala $[-1; 1]$

dice quanto "correlati" sono i valori assunti dalle due waa

ad es X, Y continue con pdf congiunta $f_{X,Y}$:



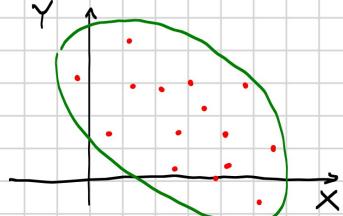
correlate positivamente

$$\rho(X, Y) \approx 0,8$$



relatione lineare esatta

$$\rho(X, Y) \approx -1$$

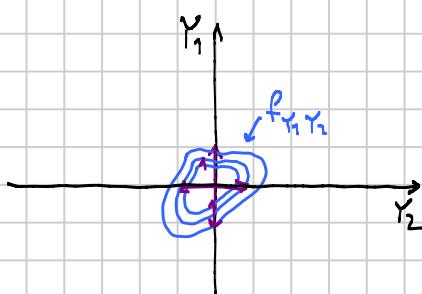
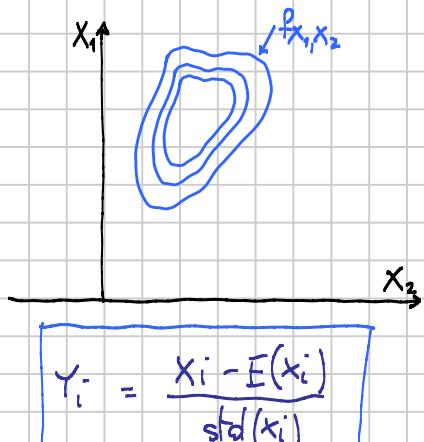


correlate negativamente

$$\rho(X, Y) \approx -0,6$$

• trasformazione che standardizza le componenti

→ applica le due trasformazioni nell'ordine



$$E(Y) = 0$$

$$C(Y) = g(Y) = p(X)$$

$$Y_i = \frac{X_i - E(X_i)}{\text{std}(X_i)}$$

$$Y = \text{diag}(\text{std}(X)^{-1})(X - E(X))$$

$$Y_{ij} = \frac{X_{ij} - \bar{X}_j}{S_{X_j}}$$

$$\left(\frac{(n, m)}{X - X.\text{mean}(\text{axis}=0, \text{keepdims=True})} \right) / X.\text{std}(\text{axis}=0 \dots)$$

■ VERSO LA PCA, PROPRIETÀ AVANZATE MATERICE DI COVARIANZA

X vettore aleatorio di dim m , $C(X) \in M_{m,m}$ la sua matrice di covarianza

1) Varianza di una somma generale

Recall : se X_1, X_2, \dots, X_m sono indip. allora

$$\text{Var}(x_1 + \dots + x_m) = \text{Var}(x_1) + \dots + \text{Var}(x_m)$$

$$\text{Var}(x_1 + \dots + x_m) = \text{Var}(e^T X) \quad \text{dove } e = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \in M_{m,1} \approx \mathbb{R}^m \quad e^T \in M_{1,m}$$

$$e^T X = e \cdot X = \sum_{i=1}^m e_i x_i = \sum_{i=1}^m x_i$$

$$Y = e^T X \quad Y \in \mathbb{R}^k \quad k = 1 \quad Y = BX \quad B \in M_{k,m} \quad \stackrel{e^T}{=} \stackrel{\mathbb{R}^m}{X} \rightarrow \stackrel{\mathbb{R}}{Y}$$

$$\begin{aligned} \text{Var}(x_1 + \dots + x_m) &= \text{Var}(Y) = C(Y) = C(BX) = BC(X)B^T = e^T C(X) e \\ &= \sum_{i,j} C(X)_{ij} = \sum_{i,j} \text{Cov}(x_i; x_j) \end{aligned}$$



$$\boxed{\text{Var}(x_1 + \dots + x_m) = \sum_{i=1}^m \text{Var}(x_i) + 2 \sum_{i < j} \text{Cov}(x_i; x_j)}$$

2) Varianza di una combinazione lineare

$$\text{Var}(\alpha_1 x_1 + \dots + \alpha_m x_m) = \text{Var}(\alpha^T X) = \dots = \alpha^T C(X) \alpha$$

→ La varianza è sempre non negativa :

$$\forall \alpha \in \mathbb{R}^m \quad \alpha^T C(X) \alpha \geq 0$$

$C(X)$ è definita non negativa e simmetrica

➡ La matrice di covarianza ha m autovalori reali non negativi

HW : è vero anche per $g(x)$!

INTRODUZIONE AL MACHINE LEARNING

Note Title

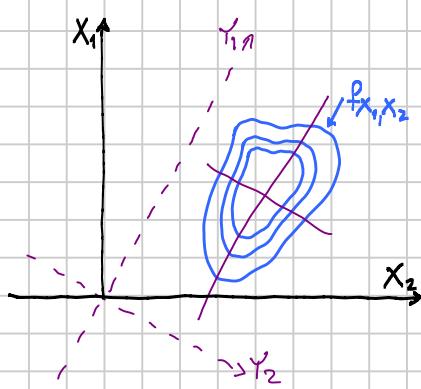
ora 10

11/03/2025

• PRINCIPAL COMPONENT ANALYSIS (PCA)

è una trasformazione lineare che scorcola le componenti (è una rotazione)

$$\Sigma = C(X) \quad C(Y) = \begin{pmatrix} ? & & 0 \\ & \ddots & \\ 0 & & ? \end{pmatrix}$$



Costruisco la matrice degli autovettori di Σ :

$$V = \begin{pmatrix} | & | & | \\ v_1 & v_2 & \dots & v_m \\ | & | & | \end{pmatrix}, \Lambda = \begin{pmatrix} \lambda_1 & & & 0 \\ & \ddots & & \\ 0 & & \ddots & \lambda_m \end{pmatrix}$$

🚩 $V^T V = V V^T = I$

questo mi dice che V è ortogonale, ovvero è una rotazione

$$[V^T V]_{ij} = \sum_k [V^T]_{ik} V_{kj} = \sum_k V_{ki} V_{kj} = \sum_k (v_i)_k (v_j)_k = v_i \cdot v_j = \begin{cases} 1 & i=j \\ 0 & \text{altri.} \end{cases} = I_{ij}$$

🚩 $V^T = V^{-1}$

$\Rightarrow V V^T = V V^{-1} = I$

□

🚩 $\Sigma V = V \Lambda$

(click)

$$[\Sigma V]_{ij} = \sum_k \sum_{ik} V_{kj} = \sum_k \sum_{ik} (v_j)_k = [\sum v_j]_i = [\lambda_j v_j]_i = \lambda_j v_{ij} = [V \Lambda]_{ij}$$

- PCA: la trasformazione lineare $Y = V^T X$ è una rotazione che scorcola le componenti: $C(Y) = \Lambda$

$$C(Y) = C(V^T X) = V^T C(X) V = V^T \Sigma V = V^T V \Lambda = \Lambda$$

□

Recall:

una matrice quadrata, simmetrica, $\Sigma \in M_n$ definita non negativa (come è $C(X)$) ammette una n autovettori $v_1, \dots, v_m \in \mathbb{R}^n$ e i corrispondenti autovalori $\lambda_1, \dots, \lambda_m \geq 0$ tali che

$$\sum v_k = \lambda_k v_k \quad k=1, 2, \dots, m$$

i v_k possono essere presi di norma 1 e sono fra loro ortogonali:

$$v_j \cdot v_k = \begin{cases} 1 & j=k \\ 0 & \text{altrimenti} \end{cases}$$

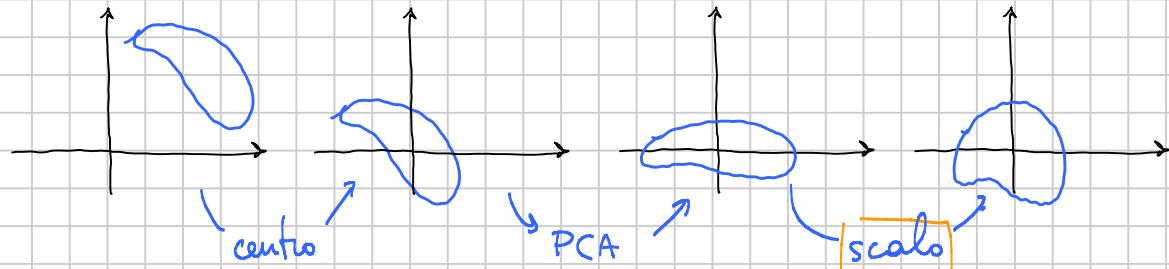
* Ci sono due approcci standard alla PCA

1) traslo nell'origine \rightarrow ruoto \rightarrow standardizzo

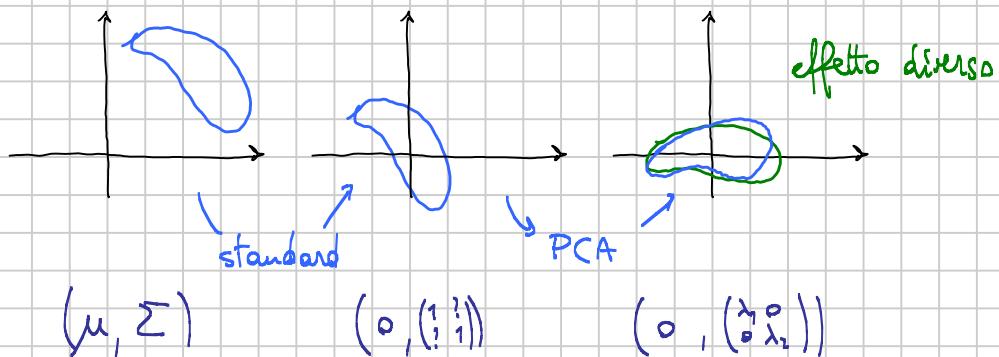
2) traslo nell'origine \rightarrow standardizzo \rightarrow ruoto (\rightarrow standardizzo)

\hookrightarrow 2) equivale a 1) ruotando con la matrice $\rho(x)$ invece di $C(x)$

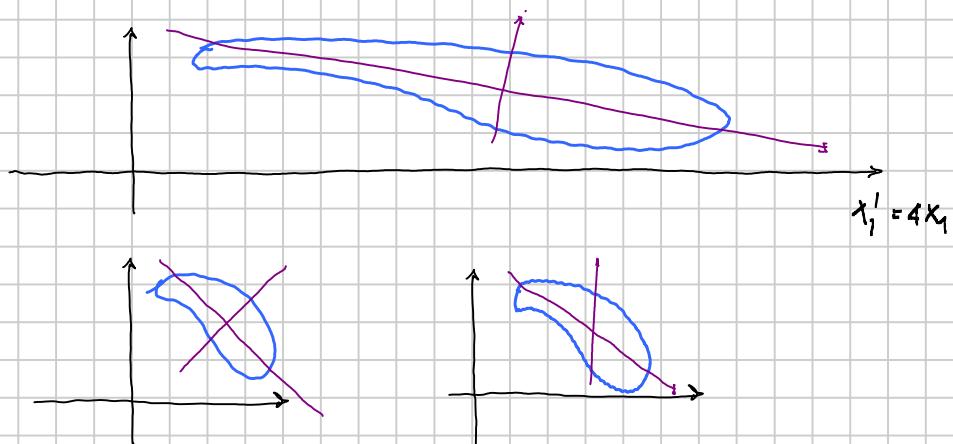
* Non è chiaro in generale cosa sia meglio



richiede che nella PCA si sia ridotto bene $n \rightarrow k$
(vedi dopo)



* Le variante originali dipendono dalle unità di misura e la matrice di covarianza ne risente



\rightarrow Se le variabili hanno unità di misura completamente diverse, conviene il metodo 2)



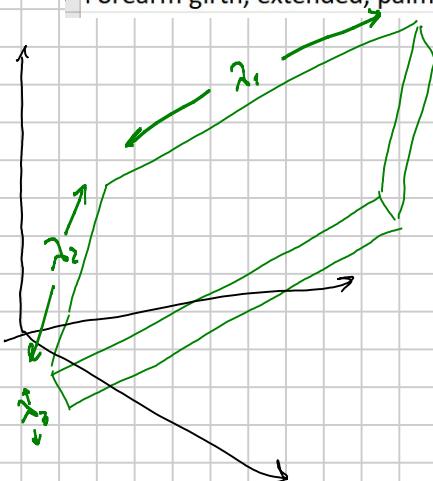
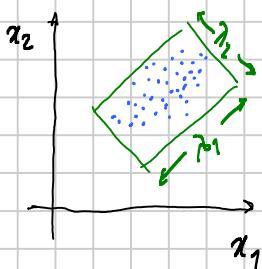
■ FACTOR ANALYSIS

→ sintetizzare le variabili in fattori che siano di meno, abbiano significato e siano fra loro indipendenti

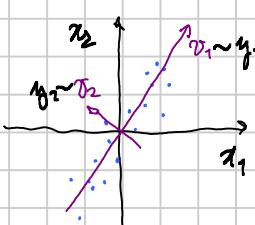
esempio : dataset body

id	x_1	x_2	\dots	x_{21}	a	w	h	g
----	-------	-------	---------	----------	-----	-----	-----	-----

- f_1 : global size
 - f_2 : width
 - f_3 : età/tasso muscolare
 - f_4 : ??? differenze sottili
- ...



* Quando m è grande, esistono molte componenti di Y con varianza piccola, "schiazzate". La analisi fattoriale, consiste nel ridurre le dimensioni ($m \rightarrow k$) senza perdere troppa informazione. La PCA è il metodo più semplice per farlo.

*  Factor loadings : $w_1 - w_k$ hanno delle direzioni nelle variabili originali

$$y_1 = 1,1 x_1 + 0,8 x_2 \quad w_1 = (1,1 ; 0,8)$$

L'analisi dei seguì dei factor loadings ci può dire qualcosa sul significato dei fattori

Biacromial diameter (see Fig. 2)

Biiliac diameter, or "pelvic breadth" (see Fig. 2)

Bitrochanteric diameter (see Fig. 2)

Chest depth between spine and sternum at nipple level,

Chest diameter at nipple level, mid-expiration

Elbow diameter, sum of two elbows

Wrist diameter, sum of two wrists

Knee diameter, sum of two knees

Ankle diameter, sum of two ankles

Shoulder girth over deltoid muscles

Chest girth, nipple line in males and just above breast

Waist girth, narrowest part of torso below the rib cage,

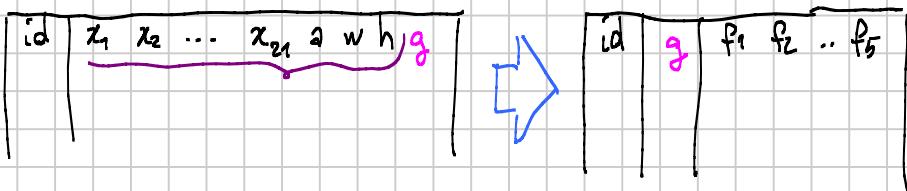
Navel (or "Abdominal") girth at umbilicus and iliac crest,

Hip girth at level of bitrochanteric diameter

Thigh girth below gluteal fold, average of right and left

Bicep girth, flexed, average of right and left girths

Forearm girth, extended, palm up, average of right and

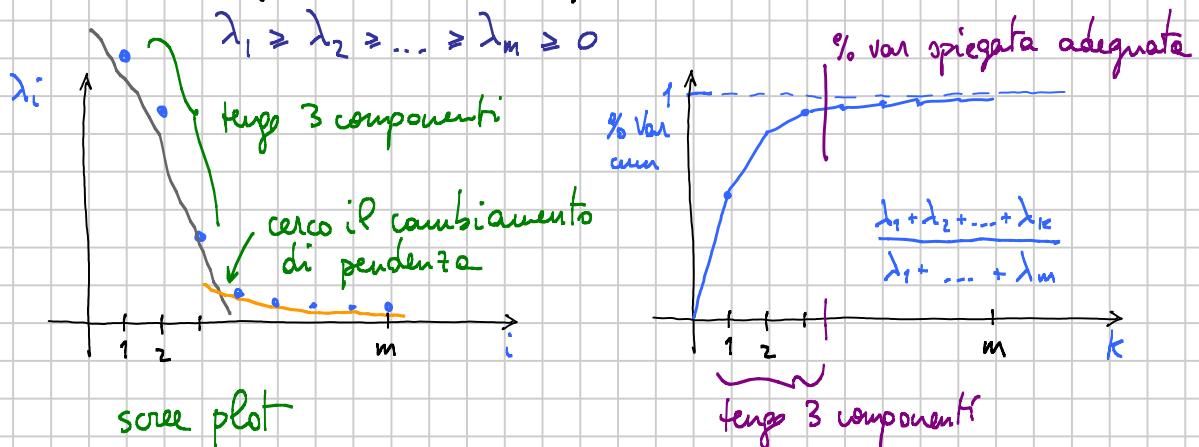


* La somma delle varianze prima e dopo è la varianza totale e si conserva

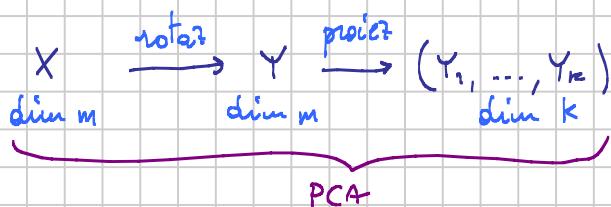
$$\text{Var}(x_1) + \dots + \text{Var}(x_m) = \lambda_1 + \dots + \lambda_m = \text{varianza totale}$$

$$\text{tr}(\Sigma) = \text{tr}(\Lambda)$$

→ tipicamente gli autovalori vengono estratti in ordine decrescente



* ciò permette di ridurre le dimensioni del vettore mantenendo la



distribuzione molto simile
e rendendo le componenti
scorrelate

→ quando si standardizza dopo la PCA è fondamentale ridurre bene

$m \rightarrow k$, perché se no si amplifica il rumore delle componenti

con λ_i piccolo

$$Y_i \rightarrow \frac{Y_i}{\sqrt{\lambda_i}}$$

→ se dopo la PCA non si standardizza, è essenzialmente solo una rotazione, perciò le tecniche che si basano solo sulle distanze fra i punti sono indifferenti alle PCA,

DISTRIBUZIONE GAUSSIANA

MULTIDIMENSIONALE

$X = (x_1, x_2, \dots, x_p)$ a valori in \mathbb{R}^p

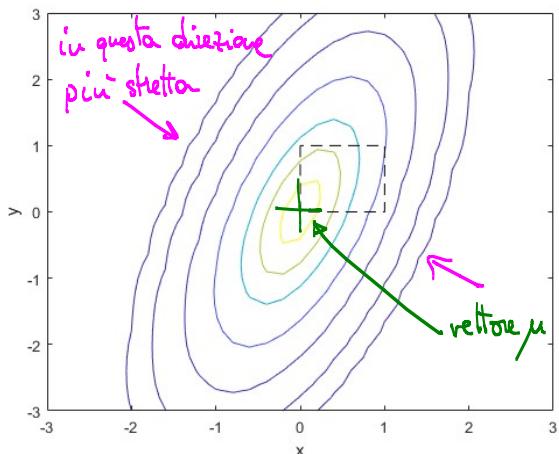
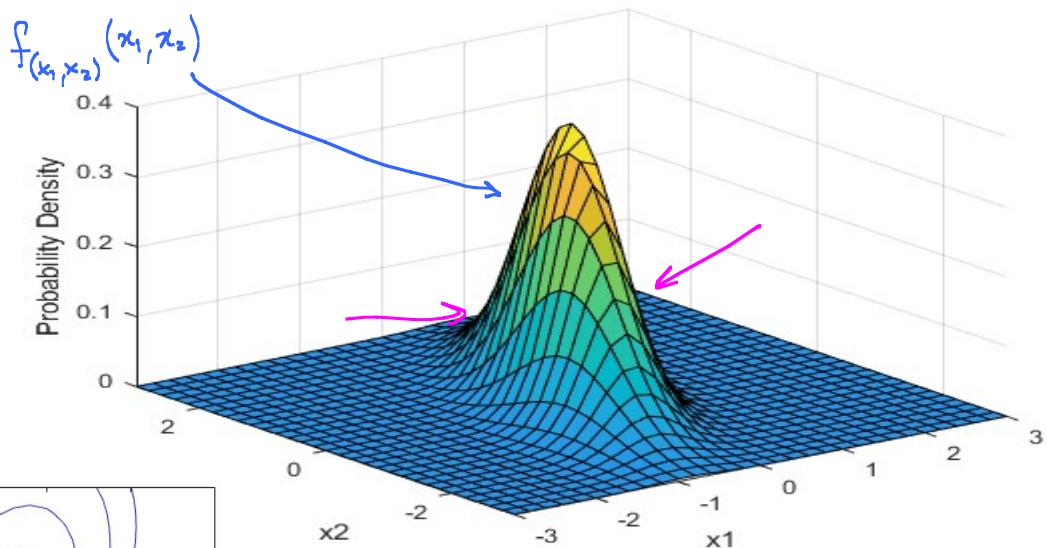
$X \sim \mathcal{N}(\mu, \Sigma)$

retta media e RP
matrice di covarianza $\in M_{p,p}$

* μ e Σ sono completamente liberi e determinano la distribuzione

<https://it.mathworks.com/help/stats/multivariate-normal-distribution.html>

* esempi di pdf
congiunta per
 $p=2$



- forma a campana potenzialmente asimmetrica
- curve di livello ellissoidali, tra loro parallele
- ma non per forza con gli assi
- centro = media

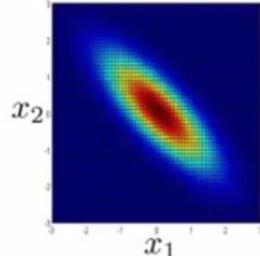
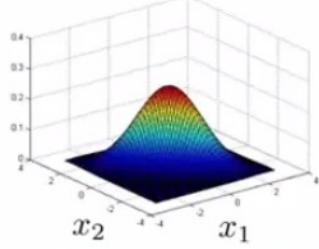
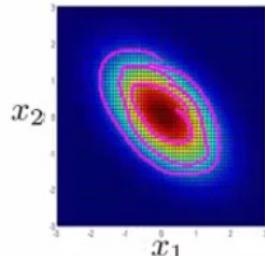
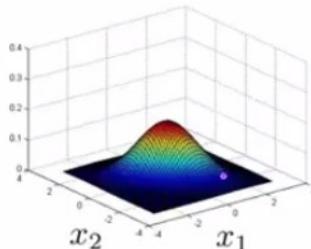
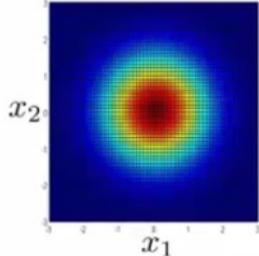
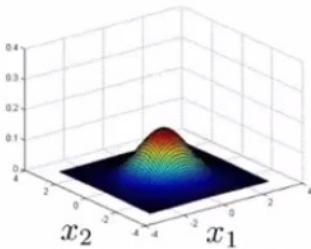
<https://www.quora.com/What-is-the-geometric-meaning-of-multivariate-Gaussian>

Multivariate Gaussian (Normal) examples

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

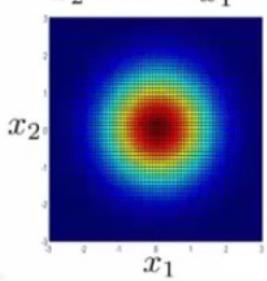
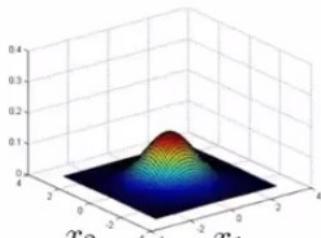
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$

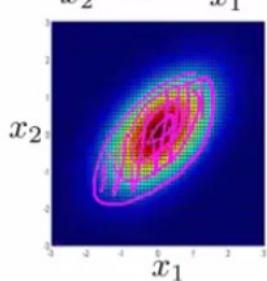
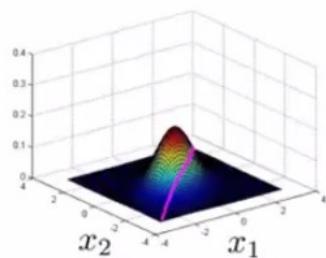


Multivariate Gaussian (Normal) examples

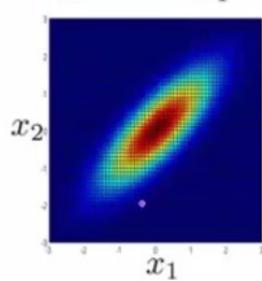
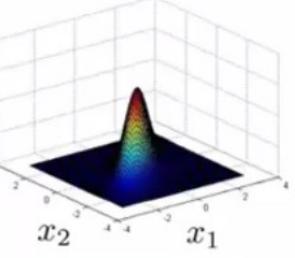
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



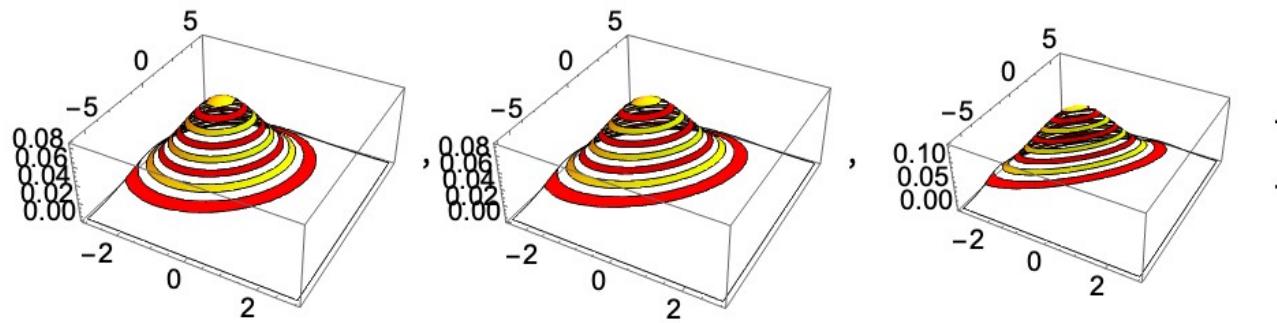
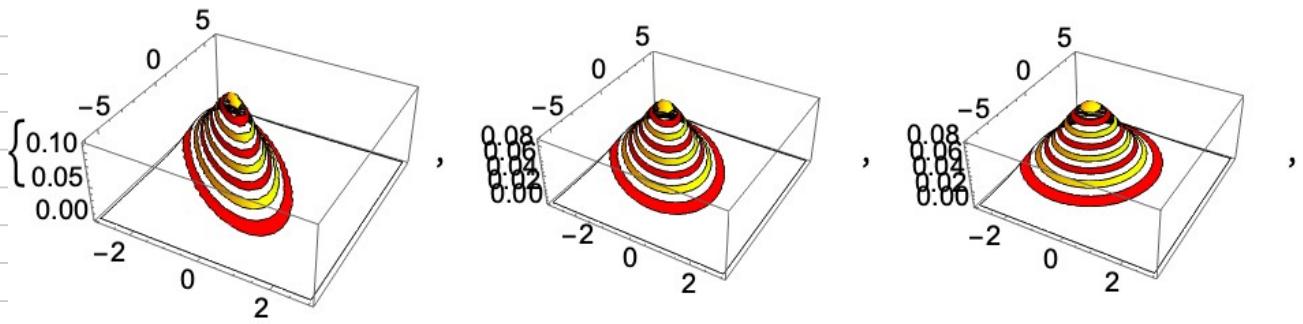
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



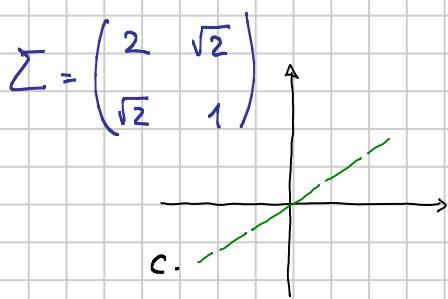
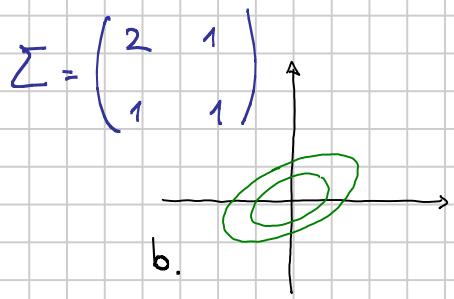
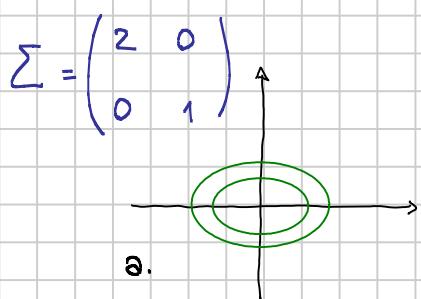
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$



<https://reference.wolfram.com/language/ref/MultinormalDistribution.html>



- solo per la Gaussiana, componenti indipendenti (\Rightarrow covarianze nulle
(di solito vale \Rightarrow e basta)
- covarianza nulla (\Rightarrow asse degli ellisoidi paralleli a quelli cartesiani)



→ se Σ è invertibile, esiste la pdf di X (altrimenti ci sono direzioni "delta")

$$f_X(x) = (2\pi)^{-\frac{p}{2}} \cdot \sqrt{\det \Sigma} \cdot \exp \left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \right\}$$

dove $\Sigma = \sum_i^{-1}$

→ l'esempio c. qui sopra è del tipo con Σ non invertibile: il vettore (x_1, x_2) giace sempre sulla retta $x_1 = \sqrt{2}x_2$

HW: verificare che se $X_2 \sim \mathcal{N}(0, 1)$, $X_1 = \sqrt{2}X_2$ e $X = (x_1, x_2)$, allora

$$\mathbb{E}(X) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ e } C(X) = \begin{pmatrix} 2 & \sqrt{2} \\ \sqrt{2} & 1 \end{pmatrix}$$

- se si proietta X sulla stessa retta, calcolando $Y = \begin{pmatrix} \frac{\sqrt{2}}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix} \cdot X$
che legge si ottiene?

- se si proietta X sulla retta ortogonale, calcolando $Z = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ -\frac{\sqrt{2}}{\sqrt{3}} \end{pmatrix} \cdot X$
che legge si ottiene?

Definizione astratta: X ha legge Gaussiana multivariata in \mathbb{R}^n se
 $\forall a \in \mathbb{R}^n$ $a \cdot X$ ha legge Gaussiana in \mathbb{R} o è
deterministica

→ se $\Sigma = I$ allora $X_1, \dots, X_p \sim \mathcal{N}(0, 1)$ i.i.d. e simmetria sferica

$$f_X(x) = (2\pi)^{-p/2} \cdot 1 \cdot \exp \left\{ -\frac{1}{2}(x-0)^T(x-0) \right\}$$

$$= C \cdot \exp \left\{ -\frac{1}{2} \|x\|^2 \right\} \text{ simmetria radiale}$$

dove $\|x\|^2 = x_1^2 + x_2^2 + \dots + x_p^2$ è il quadrato del modulo del vettore x .

→ sempre in questo caso $\|x\|^2 \sim \chi^2(p)$ (HW: check)

HW: Siano Z_1, Z_2, \dots, Z_p Gaussiane standard indipendenti. Sia $\epsilon > 0$ piccolo.

Sia $S := Z_1 + \dots + Z_p$. Sia infine $X = (x_1, x_2, \dots, x_p)$ con $x_i := Z_i - \epsilon \cdot S$.

Determinare la distribuzione di X . Discutere la correlazione fra le componenti di X al variare di ϵ .

Legge di Dirichlet

distribuzione continua : genera un vettore $X = (x_1, x_2, \dots, x_m)$ di numeri reali positivi con somma 1

$$X \sim \text{Dirichlet}(\alpha) \quad \alpha = (\alpha_1, \alpha_2, \dots, \alpha_m) \quad \alpha_i > 0$$

X rettore aleatorio di dimensione m t.c. $X_i \in [0, 1]$ $x_1 + \dots + x_m = 1$

$$f_X(x) = C_\alpha \cdot x_1^{\alpha_1-1} \cdot x_2^{\alpha_2-1} \cdots x_m^{\alpha_m-1} \quad 0 \leq x_i \leq 1 \quad x_1 + \dots + x_m = 1$$

$E(X) = \frac{1}{S} \cdot \alpha$

 $S = \alpha_1 + \alpha_2 + \dots + \alpha_m$

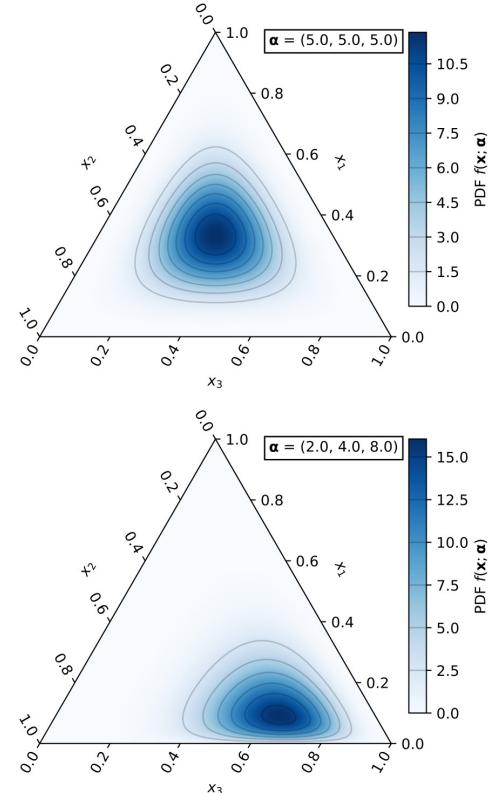
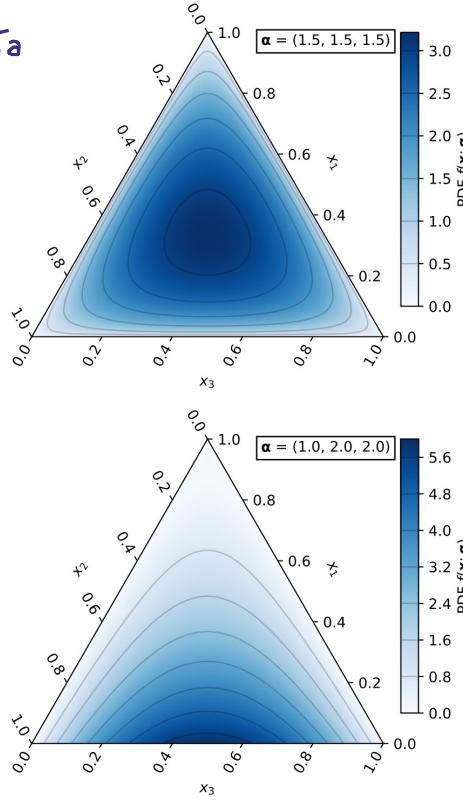
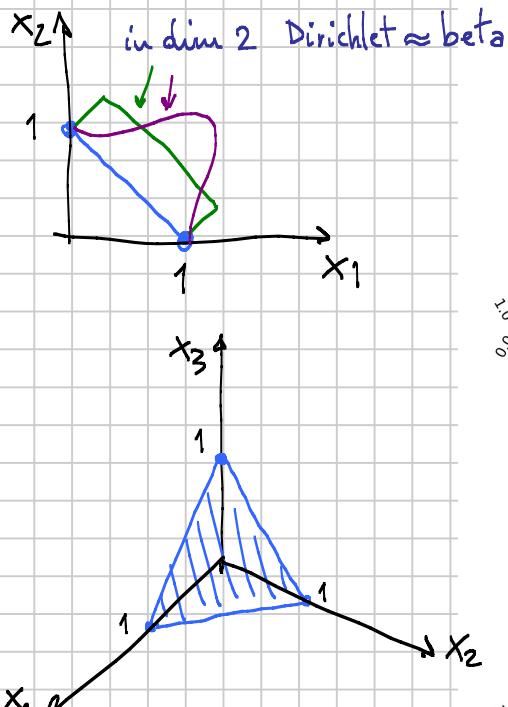
→ ad es : $m = 3 \quad \alpha_1 = 5 \quad \alpha_2 = 2 \quad \alpha_3 = 1 \quad S = 8$

allora $E(X) = \left(\frac{5}{8}, \frac{2}{8}, \frac{1}{8}\right)$

$m = 3 \quad \alpha_1 = 10 \quad \alpha_2 = 4 \quad \alpha_3 = 2 \quad S = 16$

allora $E(X) = \left(\frac{5}{8}, \frac{2}{8}, \frac{1}{8}\right)$ (uguale)

cambierà la covarianza



HW : cercare la covarianza su internet

$$m = 2 \quad X_1 \sim \text{unif}(0,1) \quad X_2 = 1 - X_1 \quad \Leftrightarrow \quad (x_1, x_2) \sim \text{Dirichlet}(1, 1)$$

Multinomiale

binomiale : prove tutte uguali; B_j con 2 esiti ciascuna : 1, 0

j	B_j	B_j^*
1	1	0
2	0	1
⋮	⋮	⋮
n	0	1
	X=12	$n-X =: \gamma$

somma : numero di insuccessi
numero di successi

$$X \sim \text{bin}(n, p) \quad p: \text{prob di 1}$$

$$Y \sim \text{bin}(n, 1-p)$$

→ il vettore $Z = (x, \gamma)$ ha componenti binomiali dipendenti

j	$B_j^{(1)}$	$B_j^{(2)}$...	$B_j^{(m)}$	j	V_j
1	0	1	...	0	1	2
2	1	0	...	0	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	0	1	...	0	2	2
	0	0	...	1	1	5
	X ₁	X ₂	...	X _m		

$$X \sim \text{multin}(n, p) \quad \text{dove}$$

$$p = (p_1, p_2, \dots, p_m) \quad \text{è un vettore di prob.}$$

$$p_1 + p_2 + \dots + p_m = 1$$

p_i : prob della categoria i

$$X_1 + X_2 + \dots + X_m = n \quad \text{vincolo}$$

$$X_i \sim \text{bin}(n, p_i)$$

* il vettore $X = (x_1, \dots, x_m)$ ha componenti binomiali dipendenti
conta il numero di esiti di ciascun tipo

$$\varphi_x(x) = P(X=x) = P(X_1=x_1, X_2=x_2, \dots, X_m=x_m)$$

$$\begin{aligned} &\text{vettore di enti possibili, tipo } x = (12, 7, 5, 0, \dots, 3) \\ &\text{e } 12+7+5+\dots+3=n \end{aligned}$$

$$= \frac{n!}{x_1! x_2! \dots x_m!} p_1^{x_1} p_2^{x_2} \dots p_m^{x_m} \quad x \in \mathbb{N}^m : x_1 + \dots + x_m = n$$

$$\text{coefficiente multinomiale} = \binom{n}{x_1, x_2, \dots, x_m} \quad \text{conta gli snugrammi}$$

$$\text{HW: Determinare } E(x) \text{ e } C(x) \quad \text{hint: } X_i = \sum_{j=1}^n H_{ij}$$

Determinare le leggi marginali di X

Q: Come genero una multinomiale?

$$\begin{array}{ccccc} \text{STATISTICA} & & & & \\ \text{T} & \text{|||} & & & 3 \\ \text{A} & \text{||} & & & 2 \\ \text{I} & \text{||} & & & 2 \\ \text{C} & \text{|} & & & 1 \\ \hline & & & & 10 \end{array}$$

$$\begin{pmatrix} 10 \\ 2, 3, 2, 2, 1 \end{pmatrix} = \frac{10!}{2! 3! 2! 2! 1!}$$

■ STIMATORI

(capitolo 6 e inizio del 7 del Ross)

Raccolto dei dati :

$X_1, X_2, X_3, \dots, X_n$ campione di dati

ad es: statura di un gruppo di persone

si immagina che i numeri raccolti: x_1, x_2, \dots, x_n siano la realizzazione di n variabili aleatorie i.i.d.: X_1, X_2, \dots, X_n

Compito della statistica inferenziale è inferire informazioni sulla legge delle X_i a partire dai dati x_1, x_2, \dots, x_n

→ esempio: misuro la pressione di un gruppo di studenti e trovo:

115 135 127 148 ... 126

→ posso almeno inferire che la pressione tipica è minore di 200

$X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$ $\mu = ?$ $\mu < 200$ molto probabilmente

Posso proporre diversi stimatori per μ , come la media campionaria $\frac{1}{n} \sum_i X_i$ o la mediana campionaria $X_{(n+1)/2}$ (nel caso Gaussiano μ è sia la media, sia la mediana, che però corrispondono a stimatori diversi e danno risultati diversi)

* In generale possono esistere diversi stimatori di una stessa grandezza, tuttavia spesso è possibile individuarne uno preferenziale.

o def Una statistica è una v.a. che è una funzione deterministica del campione $f(X_1, X_2, \dots, X_n)$

* (non vera def) Uno stimatore di un parametro Θ è una statistica $\hat{\Theta} = f(X_1, \dots, X_n)$ la cui legge è in qualche modo concentrata vicino a Θ

- def Uno stimatore consistente di un parametro θ è una famiglia di statistiche $\Theta_n \quad n=1,2,\dots$ per cui si ha che

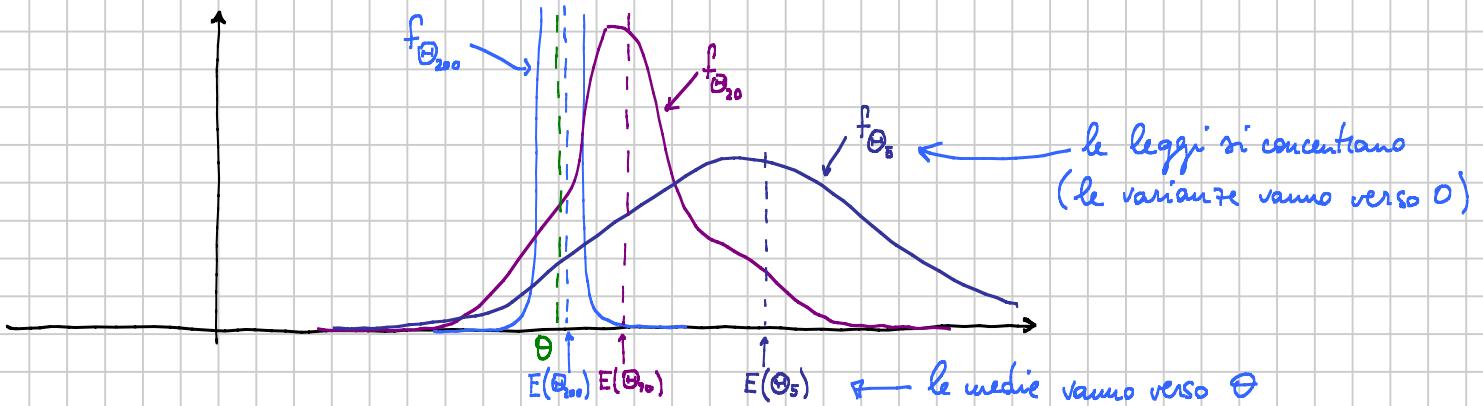
$$\Theta_n \xrightarrow[n \rightarrow \infty]{P} \theta$$

convergenza in probabilità

[vuol dire che per ogni ϵ , $P(|\Theta_n - \theta| > \epsilon) \xrightarrow{n \rightarrow \infty} 0$]

→ operativamente, basta che

$$\boxed{E(\Theta_n) \xrightarrow{n \rightarrow \infty} \theta \quad \text{e} \quad \text{Var}(\Theta_n) \xrightarrow{n \rightarrow \infty} 0}$$



→ Quando scriviamo che Θ_n dipende da n , intendiamo che al crescere del campione abbiamo un modo stabilito di aggiornare Θ usando tutti i dati a disposizione

$$\underbrace{x_1, x_2, \dots, x_n, x_{n+1}, \dots}_{X_n} \xrightarrow{\Theta_n \approx \theta}$$

- def Uno stimatore corretto di un parametro θ è una statistica Θ per cui si ha

$$E(\Theta) = \theta$$

→ altrimenti è distorto

→ si chiama bias l'errore sistematico: $E(\Theta) - \theta$

$$\text{Recall } \bar{X} = \frac{1}{n} \sum_i x_i$$

$$E(\bar{X}) = \mu \quad \mu = E(x_i)$$

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \quad \sigma^2 = \text{Var}(x_i)$$

è consistente e corretto.

• Varianza campionaria e deviazione standard campionaria

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \approx \sigma^2 \quad S_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \approx \sigma$$

* Entrambi sono stimatori consistenti di σ^2/σ , ma mentre S_x^2 è corretto, $E(S_x^2) = \sigma^2$, S_x è leggermente distorto $E(S_x) < \sigma$

→ ovviamente si intende qui che $\sigma^2 = \text{Var}(x_i)$

→ se $x_i \sim N(\mu, \sigma^2)$, si trova facilmente che $\text{Var}(S_x^2) = \frac{2\sigma^4}{n-1}$, che mostra che S_x è consistente

→ Vediamo che S_x è distorto

$$E(S_x^2) = \sigma^2 \quad E(S_x) = ?$$

$$E(S_x) \leftarrow \sigma$$

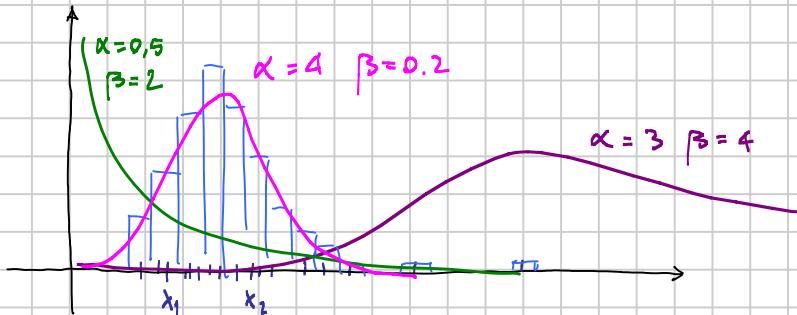
$$0 < \text{Var}(S_x) = E(S_x^2) - E(S_x)^2 \Rightarrow E(S_x)^2 < E(S_x^2) = \sigma^2$$

■ STIMATORI DI MASSIMA VEROSSIMIGLIANZA (MLE max likelihood estimators)

è un metodo per ricavare stimatori con buone proprietà anche in situazioni nuove

es 1) Campione di dati con legge gamma: vogliamo stimare i parametri

$x_1, x_2, \dots, x_n \sim \text{gamma}(\alpha, \beta)$ indipendenti x_1, x_2, \dots, x_n numeri ottenuti



Veroossimiglianza (likelihood) di una legge (α, β) , dato il campione

$(\alpha, \beta) \rightarrow \text{gamma}(\alpha, \beta) \rightarrow f_{\alpha, \beta} : \mathbb{R} \rightarrow \mathbb{R}_+$ densità

$x = (x_1, \dots, x_n)$ speriamo che abbia legge (α, β) nel senso che

$i=1, \dots, n \quad x_i \sim \text{gamma}(\alpha, \beta)$. Se è così:

$$f_x : \mathbb{R}^n \rightarrow \mathbb{R}_+ \quad f_x(t_1, \dots, t_n) = \prod_{i=1}^n f_{\alpha, \beta}(t_i)$$

densità congiunta

• def Likelihood

$$L(\alpha, \beta) := \underbrace{f_X(x_1, x_2, \dots, x_n; \alpha, \beta)}_{\text{Possati}} = \prod_{i=1}^n f_{\alpha, \beta}(x_i)$$

la scelta dei parametri α, β che massimizzano $L(\alpha, \beta)$ porta alla legge che rende più verosimili i dati osservati

→ in genere si lavora con la log-likelihood

$$\ell(\alpha, \beta) := \log L(\alpha, \beta) = \sum_{i=1}^n \log f_{\alpha, \beta}(x_i)$$

→ torna all'esempio:

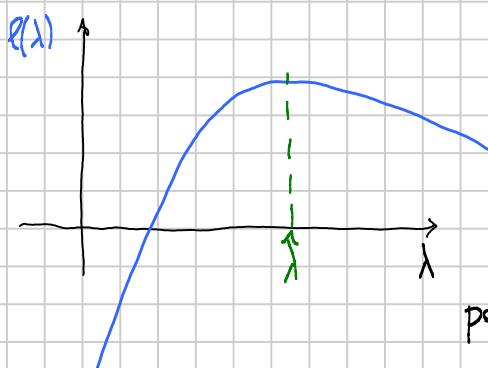
$$\begin{aligned} \ell(\alpha, \beta) &= \sum_{i=1}^n \log \left(\frac{\beta^{-\alpha}}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-x_i/\beta} \right) \\ &= -n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha-1) \sum_{i=1}^n \log x_i - \frac{1}{\beta} \sum_{i=1}^n x_i \end{aligned}$$

→ uso poi un algoritmo di ottimizzazione per trovare i migliori α, β

• Esempi classici di stimatori ML

1) $X_i \sim \text{expo}(\lambda) \quad i=1, 2, \dots, n$

$$\ell(\lambda) := \sum_{i=1}^n \log f_{\text{expo}(\lambda)}(x_i) = \sum_{i=1}^n \log (\lambda e^{-\lambda x_i}) = \sum_{i=1}^n (\log \lambda - \lambda x_i) = n \log \lambda - \lambda \sum_{i=1}^n x_i$$



è concava: derivata seconda negativa

$$\ell''(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^n x_i$$

$$\text{pongo uguale a zero } 0 = \frac{n}{\lambda} - \sum_i x_i \Leftrightarrow \hat{\lambda} = \frac{n}{\sum x_i}$$

Concludo: lo MLE di λ è $T := \frac{n}{\sum_{i=1}^n x_i} \approx \bar{x}$

$$T = (\bar{x})^{-1}$$

→ La media campionaria è MLE della media $\frac{1}{\lambda}$ nel caso esponenziale
 \bar{x} è sempre uno stimatore corretto della media, quindi l'MLE
in questo caso è corretto

$$2) X_i \sim \text{unif}(0, a)$$

$$\ell(a) = \sum_{i=1}^n \log f_{\text{unif}(0,a)}(x_i), \quad a \geq \max x_i$$

$$= \sum_{i=1}^n \log \frac{1}{a} = -n \log a, \quad a \geq \max x_i$$



per massimizzare $\ell(a)$ va preso $\hat{a} = \max x_i$

Concludo: lo MLE di a è $\hat{Y} := \max_i x_i$

* Si può dimostrare che questo stimatore è consistente ma distorto

$$E(Y) = \frac{n}{n+1} a \quad (\text{HW: verificare in modo teorico o numerico})$$

Potrei introdurre una piccola correzione: $\hat{Y}_{adj} := \frac{n+1}{n} Y = \frac{n+1}{n} \max_i x_i$
(HW: \hat{Y}_{adj} è consistente e corretto)

• NB. Lo stimatore ML è sempre consistente, e raramente corretto, ma spesso si può correggere!

$$3) X_i \sim \mathcal{N}(\mu, \sigma^2)$$

$$\ell(\mu, \sigma) = \sum_{i=1}^n \log \left(\frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} \right) = \sum_{i=1}^n \left[-\frac{1}{2} \log(2\pi) - \log \sigma - \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

$$= C - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

calcolo le derivate rispetto a μ e σ e le pongo uguali a zero

$$0 = \frac{\partial \ell(\mu, \sigma)}{\partial \mu} \Big|_{\hat{\mu}, \hat{\sigma}} = -\frac{1}{2\hat{\sigma}^2} \sum_{i=1}^n 2 \cdot (x_i - \hat{\mu}) \cdot (-1) = \frac{1}{\hat{\sigma}^2} \left[\sum_i x_i - n\hat{\mu} \right]$$

$$\Leftrightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

$$0 = \frac{\partial \ell(\mu, \sigma)}{\partial \sigma} \Big|_{\hat{\mu}, \hat{\sigma}} = -\frac{n}{\hat{\sigma}^2} - \frac{1}{2} \sum_{i=1}^n (x_i - \hat{\mu})^2 \cdot (-2) \frac{1}{\hat{\sigma}^3}$$

$$\Leftrightarrow n \hat{\sigma}^2 = \sum_i (x_i - \hat{\mu})^2 \Leftrightarrow \hat{\sigma} = \sqrt{\frac{1}{n} \sum_i (x_i - \bar{x})^2}$$

→ Nel caso Gaussiano, gli MLE di μ e σ^2/σ^2 sono \bar{X} e \tilde{S}/\tilde{S}^2

$$\bar{X} = \frac{1}{n} \sum_i x_i \quad \tilde{S} = \sqrt{\frac{1}{n} \sum_i (x_i - \bar{X})^2}$$

↑↑
 consistente e corretto consistente distorto ~ usiamo S_X che
 almeno rende S_X^2 corretto

ora 14

4) Caso Bernoulliano

$$x_i \sim \text{bin}(1, p) \quad (x_1, x_2, \dots, x_n) = (0, 1, 1, 0, 1, \dots) \quad O_1 := \sum_i x_i := \# \text{di } 1$$

$$q_{\text{bin}(1,p)}(x) = P(\text{bin}(1,p) = x) = \begin{cases} 1-p & x=0 \\ p & x=1 \end{cases} = p^x (1-p)^{1-x} \quad x=0, 1$$

$$q_{\text{bin}(n,p)}(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x=0, 1, \dots, n$$

$$\ell(p) = \sum_i \log q_{\text{bin}(1,p)}(x_i) = \sum_i [x_i \log p + (1-x_i) \log (1-p)]$$

$$= O_1 \cdot \log p + O_0 \cdot \log (1-p)$$

$$O = \ell'(\hat{p}) = \frac{O_1}{p} - \frac{O_0}{1-p} \Leftrightarrow \hat{p} O_0 = (1-\hat{p}) O_1 \Leftrightarrow \hat{p} = \frac{O_1}{O_1 + O_0} = \frac{1}{n} O_1 = \bar{x}$$

5) Caso multinomiale

$$B(i) \sim \text{multin}(1, p) \quad p = (p_1, p_2, \dots, p_m)$$

i	$B_1(i)$...	$B_m(i)$	
1	$B_1(1)$	$B_2(1)$	\dots	$B_m(1)$
2	$B_1(2)$			$B_m(2)$
:	:			
n	$B_1(n)$			$B_m(n)$

i	$B_1(i)$...	$B_m(i)$			
1	0	1	0	...	0	
2	1	0	...	0	one-hot	
:	0	0	..	1	..	0
n	0	1	..	0		

Y(i)
2
1
7
:
2

$$O_j := \sum_{i=1}^n B_j(i)$$

one-hot

$$P(B(i) = (x_1, x_2, \dots, x_m)) = p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m} = p_k \text{ t.c. } x_k = 1 \text{ e gli altri } 0$$

$$(*) \quad \ell(p_1, \dots, p_m) = \sum_i \log P(B(i) = (b_1(i), b_2(i), \dots, b_m(i)))$$

$$= \sum_i [b_1(i) \log p_1 + \dots + b_m(i) \log p_m] = \sum_{i=1}^n \sum_{j=1}^m b_j(i) \log p_j = \sum_{j=1}^m \log p_j \sum_{i=1}^n b_j(i)$$

$$= \sum_{j=1}^m O_j \log p_j$$

Questa likelihood va massimizzata sotto il vincolo che $\sum_{j=1}^m p_j = 1$

* Si trova (HW : con i moltiplicatori di Lagrange) che lo MLE è

$$p_j \approx \pi_j := \frac{1}{n} O_j$$

- Applicazioni al machine learning

→ abbiamo input x e output y

x
input

y
casuale, prevedibile almeno in parte sapendo x

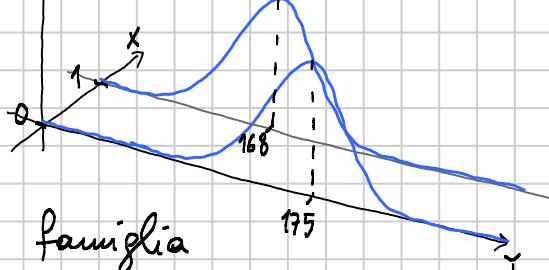
→ supponiamo che la distribuzione di y dipenda da x

ora precedente : $x, \beta \rightarrow \boxed{\quad} \rightarrow f_y$

$$x = 0 \quad Y \sim N(175, \sigma^2)$$

$$x = 1 \quad Y \sim N(168, \sigma^2)$$

adesso : $x, \beta, \gamma \rightarrow \boxed{\quad} \rightarrow f_y$



→ assumiamo ulteriormente che

la legge di y , dato x sia di una famiglia
fissata

* ad es. $Y \sim \text{multin}(1, (p_1, \dots, p_m))$

dove $p_j = p_j(x; \alpha, \beta, \gamma) \quad j = 1, 2, \dots, m$

$$(1) \quad l(\alpha, \beta, \gamma) = \sum_{i=1}^n \sum_{j=1}^m b_{j(i)} \log p_j(x_i; \alpha, \beta, \gamma)$$

$$= \sum_{i=1}^n \log p_{y(i)}(x_i; \alpha, \beta, \gamma)$$

numero di colonna dove c'è 1
nella riga i
 $y(i) \in \{1, 2, \dots, m\}$

$$\text{t.c. } b_{y(i)}(i) = 1$$

→ le funzioni $p_j()$ saranno spesso reti neurali, o modelli sofisticati
a molti parametri

→ i parametri (α, β, γ) saranno i pesi della rete, potenzialmente miliardi

* la likelihood non cambia formula, ma non si semplifica come prima;

non si calcolano le derivate, o comunque non si annullano esattamente;

si può comunque usare un ottimizzatore iterativo per trovare valori "buoni" dei parametri

- Legame con la cross-entropy loss

$$-\frac{1}{n} l(p_1, p_2, \dots) = -\frac{1}{n} \sum_{j=1}^m o_j \log p_j = -\sum_{j=1}^m q_j \log p_j = H(q, p)$$

$q_j := \frac{o_j}{n}$: "frazione di dati che hanno categoria j " $\sum_j q_j = \frac{1}{n} \sum_j o_j = \frac{1}{n} \cdot n = 1$

For discrete probability distributions p and q with the same support \mathcal{X} , this means

$$H(p, q) = - \sum_{x \in \mathcal{X}} p(x) \log q(x). \quad (\text{Eq.1})$$

→ massimizzare $l(p)$ equivale a minimizzare $H(q, p)$ al variare di p

$$-\frac{1}{n} l(\alpha, \beta, \gamma) \stackrel{(A)}{=} -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m b_j(i) \log p_j(x_i; \alpha, \beta, \gamma) = \frac{1}{n} \sum_{i=1}^n H(b(i); p(x_i; \alpha, \beta, \gamma))$$

Keras / TensorFlow

Torch

CategoricalCrossentropy class

```
keras.losses.CategoricalCrossentropy(
    from_logits=False,
    label_smoothing=0.0,
    axis=-1,
    reduction="sum_over_batch_size",
    name="categorical_crossentropy",
    dtype=None,
)
```

CrossEntropyLoss

```
CLASS torch.nn.CrossEntropyLoss(weight=None, size_average=None,
                                ignore_index=-100, reduce=None, reduction='mean',
                                label_smoothing=0.0) [SOURCE]
```

This criterion computes the cross entropy loss between input logits and target.

① Logits

es. di $p_j()$ semplici

$$p_j(x; \alpha, \beta, \gamma) \rightarrow \mathbb{R} \quad x \in \mathbb{R}^d \quad (\cancel{\alpha, \beta}) \quad w \in \mathbb{M}_{m,d} \quad b \in \mathbb{R}^m$$

$$p(x; w) \rightarrow \mathbb{R}^m$$

Logits $\rightarrow Y = b + Wx \in \mathbb{R}^m$ voglio renderli probabilità: uso il softmax $\text{sm}()$

$$\text{sm} : \mathbb{R}^m \rightarrow \mathbb{R}^m$$

$$Y \mapsto P$$

$$p(x; w) = \text{sm}(b + Wx)$$

$$\text{sm}(Y)_j = \frac{e^{Y_j}}{\sum_{i=1}^m e^{Y_i}}$$

→ aggiungo la crossentropy

$$H(q; p) = -\sum_{j=1}^m q_j \log(su(b + w \cdot x)) = -\sum_{j=1}^m q_j \log\left(\frac{e^{Y_j}}{\sum_i e^{Y_i}}\right)$$

- * $H(q; su(Y))$ è la cross-entropy calcolata dai logits: è molto più precisa, stabile e ottimizzata di fare $su()$ e $H()$ una dopo l'altra

INTRODUZIONE AL MACHINE LEARNING

Note Title

ora 16

25/03/2025

CROSS-ENTROPY LOSS

Faccio il punto

- si usa per dati categorici (= multinomiali)

i	$x(i)$	$y(i)$	\rightarrow	$B_1(i) \ B_2(i) \ B_3(i) \ B_4(i) \ \dots \ B_m(i)$
1	~	2		0 1 0 0 ... 0
2	~	1		1 0 0 0 ... 0
3	~	3		0 0 1 0 ... 0
:	~	1		1 0 0 0 ... 0
:				.
n	~	4		0 0 0 1 ... 0

m categorie $o_1 \ o_2 \ \dots \ o_m$ ← somme verticali
= osservati

- le varie categorie avranno delle probabilità $p_j = P(Y=j) \quad j=1, 2, \dots, m$

$$P(Y(i) = j) = p_j(i)$$

$$= p_1(i)^{b_1} \cdot p_2(i)^{b_2} \cdot \dots \cdot p_m(i)^{b_m} \quad \text{dove } (b_1, \dots, b_m) = \text{one-hot}(j)$$

$$\ell(\dots) = \left[\sum_{i=1}^n \sum_{j=1}^m b_j(i) \log p_j(i) \right] \quad \begin{matrix} \uparrow \\ \text{matrice 0/1 dati} \end{matrix} \quad \begin{matrix} \uparrow \\ \text{modello per le probabilità} \end{matrix}$$

- puo' esserci o no dipendenza da x

a) non c'è (dipendenza da) x : $p_j(i) = p_j$ non dipendente da i

↪ come parametri è naturale usare i p_j

$$\ell(p_1, \dots, p_m) = \sum_{i=1}^n \sum_{j=1}^m b_j(i) \log p_j = \sum_{j=1}^m \log p_j \sum_{i=1}^n b_j(i) = \sum_{j=1}^m o_j \log p_j$$

$\left. \begin{matrix} \uparrow \\ o_j \end{matrix} \right\} = -n H(\hat{p}, p)$

→ si risolve : $\hat{p}_j = \frac{o_j}{n}$ una qualche funzione

b) c'è dipendenza da x : $p_j(i) = \pi_j(x(i); \alpha, \beta, \gamma)$

$$\ell(\alpha, \beta, \gamma) = \sum_{i=1}^n \sum_{j=1}^m b_j(i) \log (\pi_j(x(i); \alpha, \beta, \gamma))$$

$\left[- \sum_{i=1}^n H(b(i), \pi(x(i); \alpha, \beta, \gamma)) \right]$

→ si cercano i parametri che massimizzano ℓ con un ottimizzatore

* Con cross-entropy loss in genere si intende $-\frac{1}{n} \ell(\dots)$

$$a) \text{loss}_{\text{CE}}(p) = H(\hat{p}, p)$$

$$b) \text{loss}_{\text{CE}}(\alpha, \beta, \gamma) = \frac{1}{n} \sum_{i=1}^n H(b(i), \pi(x(i); \alpha, \beta, \gamma))$$

→ spesso $H(q, \text{logits})$ invece che $H(q, p)$

MEAN SQUARED ERROR LOSS (MSE)

- si usa per dati Gaussiani

i	$x(i)$	$y(i)$
1	~	~
2	~	~
3	~	~
:	~	~
:	~	~
n	~	~

- distribuzione : $y(i) \sim \mathcal{N}(\mu(i), \sigma^2(i))$

$$f_{Y(i)}(z) = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma(i)} \cdot \exp \left\{ -\frac{(z-\mu(i))^2}{2\sigma^2(i)} \right\}$$

$$\ell(\dots) = \sum_{i=1}^n \left[-\log(\sqrt{2\pi}) - \log(\sigma(i)) - \frac{(y(i)-\mu(i))^2}{2\sigma^2(i)} \right]$$

ora 17

- tre casi :

$$a) \text{no dipendenza da } x \quad \mu(i) = \mu \quad \sigma^2(i) = \sigma^2$$

$$\ell(\mu, \sigma^2) = C - n \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y(i) - \mu)^2$$

$$\rightarrow \text{risolve: } \hat{\mu} = \bar{y} = \frac{1}{n} \sum_i y(i) \quad \hat{\sigma}^2 = \sqrt{\frac{1}{n} \sum_i (y(i) - \bar{y})^2}$$

b) caso omoschedastico (= varianza costante)

$$\sigma^2(i) = \sigma^2, \quad \mu(i) = \nu(x(i); \alpha, \beta, \gamma) \quad \text{una qualche funzione}$$

$$\ell(\sigma^2, \alpha, \beta, \gamma) = C - n \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y(i) - \mu(i))^2$$

$\text{SE}(y(i), \mu(i))$

$$\text{SE}(y, z) = (y - z)^2$$

$$\text{loss}_{\text{MSE}}(\alpha, \beta, \gamma) = \frac{1}{n} \sum_{i=1}^n \text{SE}(y(i), v(x(i); \alpha, \beta, \gamma))$$

$$l(\hat{\sigma}, \alpha, \beta, \gamma) = C - n \left(\log \hat{\sigma} + \frac{1}{2\hat{\sigma}^2} \text{loss}_{\text{MSE}}(\alpha, \beta, \gamma) \right)$$

$$\rightarrow \hat{\sigma} = \sqrt{\text{loss}_{\text{MSE}}(\alpha, \beta, \gamma)}$$

→ gli altri parametri si trovano minimizzando loss_{MSE}

c) tutto dipende da x

$$\mu(i) = v(x(i); \alpha, \beta, \gamma) \quad \hat{\sigma}(i) = \text{se}(x(i); \alpha, \beta, \gamma)$$

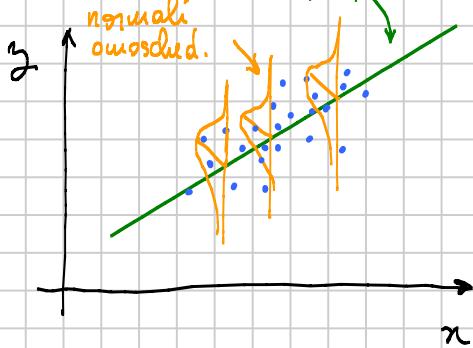
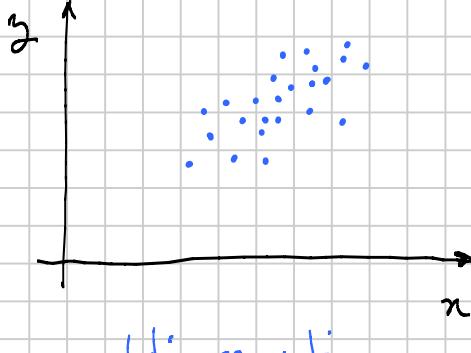
→ ottimizzazione iterativa.

REGRESSIONE

REGRESSIONE LINEARE SEMPLICE

Campione bivariato $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

"risposta media": $\mu = \beta_0 + \beta_1 x$



id	x_i	y_i
1	x_1	y_1
2	x_2	y_2
\vdots	\vdots	\vdots
n	x_n	y_n

i. x_1, x_2, \dots, x_n
deterministiche

ii. y_1, y_2, \dots, y_n
casuali

iii. $y_i \sim N(\mu_i, \sigma^2)$

iv. $\mu_i = \beta_0 + \beta_1 x_i$

* Parametri del modello: $\beta_0, \beta_1, \sigma^2$

Modello che si adotta è molto rigido:

x_i : deterministiche, esatte, assegnate

y_i : realizzazione di y_i v.a. Gaussiane indipendenti

$$Y_i \sim N(\beta_0 + \beta_1 x_i; \sigma^2)$$

not. modello: $Y = \beta_0 + \beta_1 x + e$

$e \sim N(0, \sigma^2)$

di solito ipotesi di omoschedasticità: σ^2 non dipende da i

* Si scelgono di solito x e y in modo che x sia "causa"
e y "effetto" (ad esempio con statura e peso)

è un po' più ragionevole prendere x la statura)

→ va anche tenuto conto che l'inferenza lavora bene su y

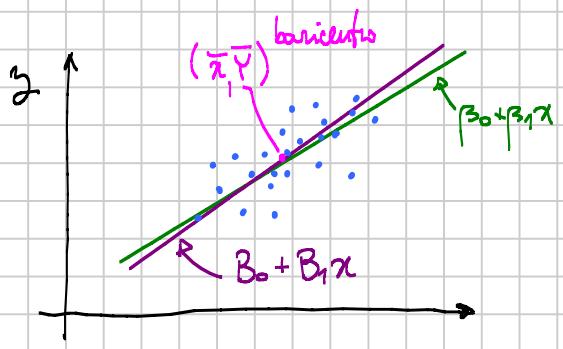
→ ci sono tre parametri incogniti: $(\beta_0, \beta_1, \sigma^2)$

→ i loro stimatori di ML si denotano: (B_0, B_1, S_e^2)

$$B_1 := \frac{\bar{x}Y - \bar{Y}\bar{x}}{\bar{x}^2 - \bar{x}^2}$$

$$B_0 := \bar{Y} - B_1 \bar{x}$$

$$S_e := \sqrt{\frac{1}{n-2} \sum_{i=1}^n (Y_i - B_0 - B_1 x_i)^2}$$



retta stimata: miglior fit dei punti

$$* \bar{Y} = \beta_0 + \beta_1 \bar{x}$$

$$* \bar{x}\bar{Y} = \frac{1}{n} \sum_{i=1}^n x_i Y_i \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

covarianza campionaria

$$B_1 = \frac{\bar{x}\bar{Y} - \bar{x}\bar{Y}}{\bar{x}^2 - \bar{x}^2} \stackrel{(check)}{=} \frac{S\text{Cov}(x, Y)}{S_x^2} = \frac{\frac{1}{n-1} \left(\sum_i x_i Y_i - n \bar{x} \bar{Y} \right)}{\frac{1}{n-1} \left(\sum_i x_i^2 - n \bar{x}^2 \right)} = S\text{Corr}(x, Y) \cdot \frac{S_y}{S_x}$$

* MLE per σ^2 avrebbe $\frac{1}{n}$ invece di $\frac{1}{n-2}$, ma S_e^2 è corretto così

① Come trovare β_0 e β_1

$$\text{loss}_{MSE}(\beta_0, \beta_1) := \frac{1}{n} \sum_{i=1}^n (Y_i - v(x_i; \beta_0, \beta_1))^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2$$

→ risolvo (cerco il minimum) facendo le derivate:

$$0 \stackrel{!}{=} \frac{\partial \text{loss}}{\partial \beta_0} = \frac{1}{n} \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot 2 \cdot (-1)$$

$$\Leftrightarrow 0 = \frac{1}{n} \sum_i Y_i - \hat{\beta}_0 - \hat{\beta}_1 \frac{1}{n} \sum_i x_i = \bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}$$

$$0 \stackrel{!}{=} \frac{\partial \text{loss}}{\partial \beta_1} = \frac{1}{n} \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot 2 \cdot (-x_i)$$

$$\Leftrightarrow 0 = \frac{1}{n} \sum_i x_i Y_i - \hat{\beta}_0 \frac{1}{n} \sum_i x_i - \hat{\beta}_1 \frac{1}{n} \sum_i x_i^2 = \bar{x}\bar{Y} - \hat{\beta}_0 \bar{x} - \hat{\beta}_1 \bar{x}^2$$

sistema

→ risolvo il sistema e trovo le formule sopra

ora 18

* Infine trovo S_e

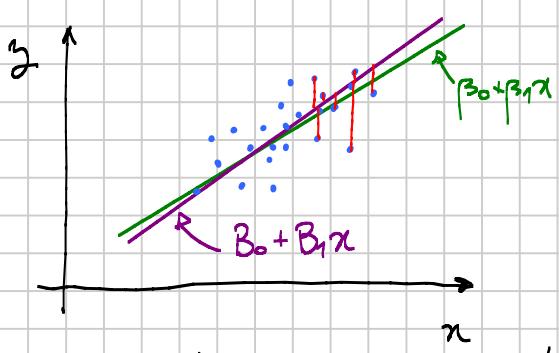
$$\text{MLE}(\hat{\sigma}) = \sqrt{\text{loss}_{MSE}(\beta_0, \beta_1)} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2} =: \sqrt{\frac{1}{n} \text{SSR}}$$

somma dei quadrati dei residui

→ tuttavia, vediamo che $E(\text{SSR}) = (n-2)\hat{\sigma}^2$ quindi preferiamo usare

$$S_e = \sqrt{\frac{1}{n-2} \text{SSR}}$$

$$E(S_e^2) = \hat{\sigma}^2 \quad \text{stimatore corretto}$$



residui
 $R_i := Y_i - B_0 - B_1 x_i$
 errori
 $e_i := Y_i - \hat{Y}_i \sim N(0, \sigma^2)$

* La retta di regressione stimata $\hat{Y}_i = \hat{B}_0 + \hat{B}_1 x_i$ è quella che minimizza $\sum_i R_i^2$

■ Teorema di Cochran

(versione ML)

Supponiamo di essere nel caso MSE loss omoschedastico :

$x(i)$ (eventualmente vettori)

$$Y(i) \sim N(\mu(i), \sigma^2(i)^2) \quad \sigma^2(i) \equiv \sigma^2 \text{ omosch.}$$

$$\begin{aligned} \mu(i) &= \gamma(x(i); \gamma_1, \dots, \gamma_k) \stackrel{!}{=} \gamma_1 \cdot c_1(x(i)) + \gamma_2 \cdot c_2(x(i)) + \dots + \gamma_k \cdot c_k(x(i)) \\ &= \gamma \cdot c(x(i)) \quad \text{dipendenza lineare dai par.} \end{aligned}$$

Come già detto, gli stimatori MLE dei parametri si trovano minimizzando

$$\text{loss}_{\text{MSE}}(\gamma_1, \dots, \gamma_k) = \frac{1}{n} \sum_{i=1}^n (Y(i) - \gamma \cdot c(x(i)))^2$$

Ovvero minimizzando $w(\gamma) = \|Y - \gamma \cdot c(x)\|^2$ norma di \mathbb{R}^n

Al variare di γ $\gamma \cdot c(x)$ è un sottospazio vettoriale di \mathbb{R}^n di dimensione k

Si trova che $\hat{\gamma}$, lo stimatore MLE, è la proiezione ortogonale di Y su questo sottospazio.

Di conseguenza $SS_R = w(\hat{\gamma})$ è la somma dei quadrati dei residui

Allora : 1) $\hat{\gamma}$ è uno stimatore corretto di γ

2) $\frac{SS_R}{\sigma^2} \sim \chi^2(n-k) \Rightarrow \frac{SS_R}{n-k}$ è uno stimatore corretto di σ^2

3) $\hat{\gamma}$ e SS_R sono variabili aleatorie indipendenti

Thm (Cochran) : enunciato per le applicazioni

$$X_1, X_2, \dots, X_n \quad X_i \sim \mathcal{N}(\mu_i, \sigma^2) \text{ indipendenti}$$

\uparrow dati "omoschedastici"

supponiamo di sapere che $\mu = (\mu_1, \dots, \mu_n) \in V \subseteq \mathbb{R}^n$

per un qualche ss.v. V di \mathbb{R}^n assegnato, con $k = \dim(V)$

Allora :

1) Lo stimatore ML di μ è la proiezione ortogonale $\pi_V(x)$

↳ in particolare può essere trovato minimizzando $\|x - y\|$ per $y \in V$.

2) $\pi_V(x)$ è uno stimatore corretto

3) $W := \|x - \pi_V(x)\|^2$ è indipendente da $\pi_V(x)$ e $\boxed{\frac{W}{\sigma^2} \sim \chi^2(n-k)}$

Dim Premette : sia v_1, \dots, v_k base ortonormale di V che estendo
a base di \mathbb{R}^n $v_1, \dots, v_k, v_{k+1}, \dots, v_n$

$$\forall x \in \mathbb{R}^n \quad \pi_V(x) = \sum_{i=1}^k v_i \cdot x \ v_i$$

$$x = \sum_{i=1}^n v_i \cdot x \ v_i$$

$$x - \pi_V(x) = \sum_{i=k+1}^n v_i \cdot x \ v_i$$

1) likelihood $L(\mu) = f_x(x_1, x_2, \dots, x_n; \mu) = C \cdot \exp\left\{-\frac{1}{2}(x-\mu)^T Q(x-\mu)\right\} = C \cdot \exp\left\{-\frac{1}{2\sigma^2} \|x-\mu\|^2\right\}$

$$Q = C(x)^{-1} = (\sigma^2 I)^{-1} = \frac{1}{\sigma^2} I$$

$$\hat{\mu} = \arg \max_{\mu \in V} L(\mu) = \arg \min_{\mu \in V} \|x - \mu\|^2$$

riccome minimizza la distanza, è uguale a

$$\boxed{\hat{\mu} = \pi_V(x)} \quad \diamond$$

2) $\pi_V(x)$ è corretto?

$$E(\pi_V(x)) = E\left(\sum_{i=1}^k v_i \cdot x \ v_i\right) = \sum_{i=1}^k v_i \cdot E(x) v_i \stackrel{\substack{\mu = (x_1, \dots, x_n) \\ \| \cdot \|}}{=} \pi_V(\mu) = \mu$$

◊

3) Ruotiamo il riferimento cartesiano in modo da usare le componenti v_1, \dots, v_n

$$N = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

matrice ortogonale $N^T N = N N^T = I$ $N^T = N^{-1}$

è una rotazione di \mathbb{R}^n $\|Nx\| = \|x\|$ (check)

$$N v_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ 0 \end{pmatrix} = e_i \Rightarrow \left\{ \begin{array}{l} N \pi_v(x) = N \sum_{i=1}^k v_i \cdot x \ v_i = \sum_i v_i \cdot x \ N v_i = \sum_1^k v_i \cdot x \ e_i \\ N x = N \sum_{i=1}^n v_i \cdot x \ v_i \\ N(x - \pi_v(x)) = N \sum_{i=k+1}^n v_i \cdot x \ v_i \end{array} \right.$$

e_i k
 $= \sum_1^n v_i \cdot x \ e_i$
 $= \sum_{k+1}^n v_i \cdot x \ e_i$

$$N \pi_v(x) = (v_1 \cdot x, v_2 \cdot x, \dots, v_k \cdot x, 0, 0, \dots, 0)$$

$$N(x - \pi_v(x)) = (0, 0, \dots, 0, v_{k+1} \cdot x, \dots, v_n \cdot x)$$

↳ in particolare, riccoce $\mu \in V$, $N\mu = N\pi_v(\mu) = (v_1 \cdot \mu, \dots, v_k \cdot \mu, \underbrace{0, \dots, 0}_I)$

$$X \sim \mathcal{N}(\mu, \sigma^2 I)$$

$$Z := NX \sim \mathcal{N}(N\mu, \sigma^2 NIN^T)$$

ha componenti indip.

$$\begin{aligned} & \text{indipendenti} \quad (Z_1, \dots, Z_k, 0, \dots, 0) = N\pi_v(x) \quad \rightarrow \quad N^T N \pi_v(x) = \pi_v(x) \quad \text{indipendenti} \\ & \quad (0, \dots, 0, \underbrace{Z_{k+1}, \dots, Z_n}_{\text{tutte di media } 0}) = N(x - \pi_v(x)) \quad \rightarrow \quad N^T N(x - \pi_v(x)) = x - \pi_v(x) \end{aligned}$$

$$\text{Infine } W := \|x - \pi_v(x)\|^2 = \|N(x - \pi_v(x))\|^2 = \sum_{k+1}^n Z_i^2 \quad Z_i \sim \mathcal{N}(0, \sigma^2)$$

\uparrow
N rotat

$$\text{Quindi } \frac{W}{\sigma^2} = \sum_{k+1}^n \left(\frac{Z_i}{\sigma} \right)^2 \sim \chi^2(n-k)$$

□

INTRODUZIONE AL MACHINE LEARNING

Note Title

ora 19

08/04/2025

• Ancora sul thru di Cochran

hp : $i=1,2,\dots,n$ $x(i), Y(i)$ $x(i)$ qualsiasi, $Y(i)$ numeri reali

$Y(i) \sim N(\mu(i), \sigma^2)$ σ^2 incognita

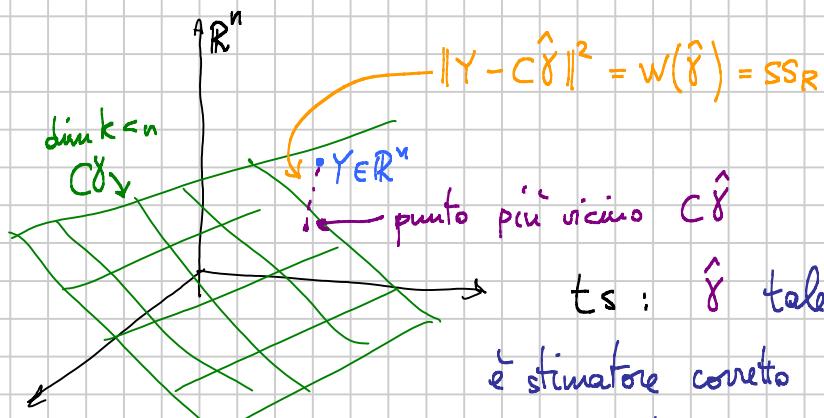
$\mu(i) = \sum_{j=1}^k \gamma_j \cdot c_j(x(i))$ k parametri, le medie dipendono come si vede dalle $x(i)$ e linearmente dai parametri

$\rightarrow C \in M_{n,k}$ matrice $c_{ij} := c_j(x(i))$

$\rightarrow \mu \in \mathbb{R}^k$ vettore $\mu_i := \mu(i)$

$\rightarrow \mu = C\gamma$ $\gamma \in \mathbb{R}^k$

* μ in funzione di γ spazia su un sottospazio di \mathbb{R}^n con dimensione k



ts : $\hat{\gamma}$ tale che $C\hat{\gamma}$ è la proiezione di Y su $C\gamma$
 è stimatore corretto di γ ; indipendente da SS_R ;
 $\frac{SS_R}{n-k}$ è stimatore corretto di σ^2 , $\frac{SS_R}{\sigma^2} \sim \chi^2(n-k)$

* Siccome è tutto lineare, c'è una formula per la proiezione :



$$\boxed{\hat{\gamma} = (C^T C)^{-1} C^T Y}$$

Verifico : scrivo $w(\gamma)$ e minimizzo facendo le derivate

$$w(\gamma) = \|Y - C\gamma\|^2 = \sum_{i=1}^n (Y_i - (C\gamma)_i)^2 = \sum_{i=1}^n (Y_i - \sum_{j=1}^k c_{ij} \gamma_j)^2$$

$$\partial \frac{\partial w(\gamma)}{\partial \gamma_h} = \sum_{i=1}^n 2(Y_i - \sum_{j=1}^k c_{ij} \gamma_j) \cdot (-c_{ih}) \Leftrightarrow \sum_i \sum_j c_{ij} \hat{\gamma}_j c_{ih} = \sum_i Y_i c_{ih} \quad \forall h = 1, \dots, k$$

$$(C^T C)_{hh} = \sum_{ij} c_{hi}^T c_{ij} \hat{\gamma}_j = \sum_i c_{hi}^T Y_i = (C^T Y)_h \Leftrightarrow C^T C \hat{\gamma} = C^T Y \Leftrightarrow \hat{\gamma} = (C^T C)^{-1} C^T Y$$

* Esempio : campione Gaussiano $Y_1, \dots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$

$$\mu(i) = \mu = \underbrace{\gamma_1}_{c_1(\alpha(i))} \cdot \underbrace{1}_{k=1} \quad C = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$C^T C = (1 \cdots 1) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = n \quad C^T Y = (1 \cdots 1) \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \sum_{i=1}^n Y_i$$

$$\hat{\mu} = \bar{Y}_1 = (C^T C)^{-1} C^T Y = \frac{1}{n} \sum_{i=1}^n Y_i = \bar{Y}$$

$$SS_R = \sum_i (Y_i - \mu(i))^2 = \sum_i (Y_i - \bar{Y})^2$$

$$\frac{SS_R}{n-1} = \frac{1}{n-1} \sum_i (Y_i - \bar{Y})^2 =: S_x^2$$

S_x^2, \bar{Y} indipendenti

$$\frac{SS_R}{S_x^2} = \frac{S_x^2}{\sigma^2} (n-1) \sim \chi^2(n-1)$$

definizione operativa t - Student : $Z \sim \mathcal{N}(0,1)$, $W \sim \chi^2(k)$ indipendenti

$$\rightarrow Z \cdot \left(\frac{W}{k} \right)^{-\frac{1}{2}} \sim t(k)$$

$$\bar{Y} \sim \mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right) \Rightarrow \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0,1)$$

$$Z \left(\frac{W}{k} \right)^{-\frac{1}{2}} = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \cdot \left(\frac{S_x^2}{\sigma^2} (n-1) / (n-1) \right)^{-\frac{1}{2}} = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \cdot \frac{\sigma^2}{S_x} = \frac{\bar{Y} - \mu}{S_x} \sqrt{n} \sim t(n-1)$$

ora 20

Distribuzione di $\hat{\gamma}$

$$\hat{\gamma} = \underbrace{(C^T C)^{-1} C^T Y}_N \quad N \text{ matrice } k \times n \quad Y \text{ vettore aleatorio Gaussiano di } \mathbb{R}^n$$

$$Y(i) \sim \mathcal{N}(\mu(i); \sigma^2), \text{ indip per } i=1, 2, \dots, n$$

$$\downarrow \text{identità } n \times n \quad Y \sim \mathcal{N}(\mu; \sigma^2 I) \Rightarrow \hat{\gamma} = NY \sim \mathcal{N}(N\mu; N\sigma^2 I N^T)$$

$$N\mu = (C^T C)^{-1} C^T C \hat{\gamma} = \boxed{\hat{\gamma}}$$

$$N\sigma^2 I N^T = \sigma^2 N I N^T = \sigma^2 N N^T = \sigma^2 (C^T C)^{-1} C^T \underbrace{[(C^T C)^{-1} C^T]^T}_{\text{imm.}} = \boxed{(C^T C)^{-1}}$$

$$= \sigma^2 (C^T C)^{-1} C^T C (C^T C)^{-1} = \sigma^2 (C^T C)^{-1}$$

rimu. \rightarrow anche $()^{-1}$ lo è

$$\hat{\gamma} \sim \mathcal{N}(\gamma; \sigma^2 (C^T C)^{-1})$$



$\hat{\gamma}$ è quindi corretto

* Esempio : regressione lineare semplice

$$i = 1, 2, \dots, n \quad x(i) \in \mathbb{R} \quad Y(i) \sim \mathcal{N}(\mu(i), \sigma^2) \quad \mu(i) = \beta_0 + \beta_1 x(i)$$

$$\mu(i) = \beta_0 + \beta_1 x(i) \quad k=2 \quad C = \begin{pmatrix} 1 & x(1) \\ & x(2) \\ & \vdots \\ & x(n) \end{pmatrix}$$

$$C^T Y = \begin{pmatrix} 1 & & & 1 \\ & x(1) & \cdots & x(n) \end{pmatrix} \begin{pmatrix} Y(1) \\ | \\ Y(n) \end{pmatrix} = \begin{pmatrix} \sum_i Y(i) \\ \sum_i x(i) Y(i) \end{pmatrix} = n \begin{pmatrix} \bar{Y} \\ \bar{x} \bar{Y} \end{pmatrix}$$

$$C^T C = \begin{pmatrix} 1 & & & 1 \\ & x(1) & \cdots & x(n) \end{pmatrix} \begin{pmatrix} 1 & x(1) \\ & \vdots \\ & x(n) \end{pmatrix} = \begin{pmatrix} n & \sum_i x(i) \\ \sum_i x(i) & \sum_i x(i)^2 \end{pmatrix} = n \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \bar{x}^2 \end{pmatrix}$$

$$\det(C^T C) = n^2 (\bar{x}^2 - \bar{x}^2) \quad (C^T C)^{-1} = \frac{1}{n} \begin{pmatrix} \bar{x}^2 & -\bar{x} \bar{x} \\ -\bar{x} \bar{x} & \frac{1}{\bar{x}^2 - \bar{x}^2} \end{pmatrix}$$

$$\hat{\gamma} = (C^T C)^{-1} C^T Y = \begin{pmatrix} \bar{x}^2 \bar{Y} - \bar{x} \bar{x} \bar{Y} \\ \bar{x}^2 - \bar{x}^2 \\ \bar{x} \bar{Y} - \bar{x} \bar{Y} \\ \bar{x}^2 - \bar{x}^2 \end{pmatrix} = : \begin{pmatrix} B_0 \\ B_1 \end{pmatrix}$$

HW: verificare che B_0 forni

$$B_0 = \bar{Y} - B_1 \bar{x}$$

$$\hat{\gamma} = \begin{pmatrix} B_0 \\ B_1 \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} ; \frac{\sigma^2}{n(\bar{x}^2 - \bar{x}^2)} \begin{pmatrix} \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \right)$$

$$SS_R = \sum_i (Y(i) - (B_0 + B_1 x(i)))^2$$

$$S_e^2 := \frac{SS_R}{n-2} \approx \sigma^2 \text{ indip da } \hat{\gamma}$$

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{(n-2)}$$

• Inferenza : test di regressione

🚩 Q: Y dipende davvero da x o no?

Parametro : β_1 . Ipotesi :

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Stimatore B_1 , si trova la sua distribuzione : $B_1 \sim \mathcal{N}(\beta_1 ; \frac{\sigma^2}{n(\bar{x}^2 - \bar{x}^2)})$,

si ricava la funz. ancillare

f. ancillare

$$\frac{B_1 - \beta_1}{\sigma} \sqrt{n(\bar{x}^2 - \bar{x}^2)} \sim \mathcal{N}(0, 1) \Leftrightarrow \frac{B_1 - \beta_1}{S_e} \sqrt{n(\bar{x}^2 - \bar{x}^2)} \sim t(n-2) \quad (\text{check})$$

$$\rightarrow \text{stat} : \boxed{T := \frac{B_1}{S_e} \sqrt{n(\bar{x}^2 - \bar{x}^2)}} \stackrel{H_0}{\sim} t(n-2)$$

$$\text{Pendo } \bar{x}, \text{ quantili} : \quad RA_T = \pm q, q = F_{t(n-2)}^{-1} \left(1 - \frac{\alpha}{2} \right)$$

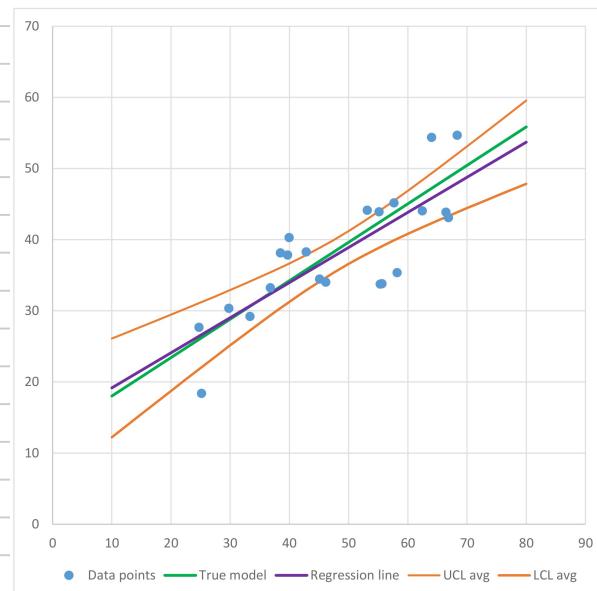
HW: scrivere p-value

Se il test dice H₀, posso ignorare le x_i e analizzare Y_1, \dots, Y_n come campione Gaussiano classico $Y_i \sim N(\mu, \sigma^2)$ $\mu = \beta_0$

ora 21

- inferenza: intervallo di confidenza per Y medio (aka per risposta media)

regressione: $E(Y)$ non è un numero $E(Y) = \beta_0 + \beta_1 x$ è una funzione



Lo stimatore puntuale è la retta di regressione: $B_0 + B_1 x$

Quello che cerco è un "intervallo" tubolare di confidenza attorno alla retta di regressione che contenga con elevata confidenza la retta della media di Y

Valore incognito da stimare: $\beta_0 + \beta_1 x$ (al variare di x)

Stimatore: $B_0 + B_1 x$. valore previsto

Distribuzione: $B_0 + B_1 x \sim N(?, ?)$

$$E(B_0 + B_1 x) = E(B_0) + x E(B_1) = \beta_0 + \beta_1 x \quad \text{corretto}$$

$$\text{Var}(B_0 + B_1 x) = \text{Cov}(B_0 + B_1 x; B_0 + B_1 x) = \text{Var}(B_0) + 2x \text{Cov}(B_0; B_1) + x^2 \text{Var}(B_1)$$

$$\begin{aligned} &= \frac{\sigma^2}{n(\bar{x}^2 - \bar{x}^2)} \left[\bar{x}^2 - 2x\bar{x} + x^2 \right] = \frac{\sigma^2}{n(\bar{x}^2 - \bar{x}^2)} \left[\bar{x}^2 - \bar{x}^2 + (x - \bar{x})^2 \right] \\ &= \frac{\sigma^2}{n} \left[1 + \frac{(x - \bar{x})^2}{\bar{x}^2 - \bar{x}^2} \right] \end{aligned}$$

consistente

... funz. ancillare ... quantili ...

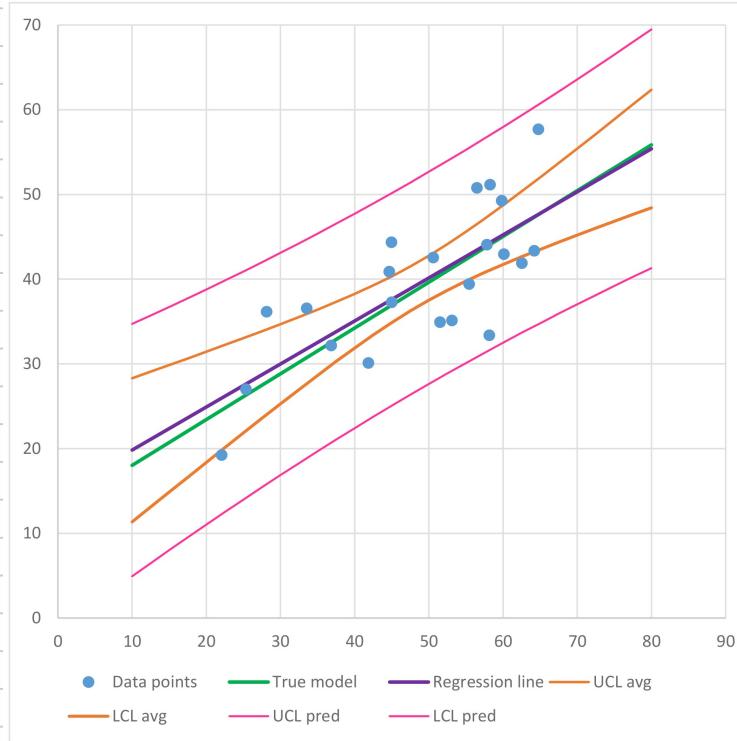
HW: completare

$$\beta_0 + \beta_1 x \in B_0 + B_1 x \pm q S_{\text{e}} \sqrt{\frac{1}{n} \left[1 + \frac{(x - \bar{x})^2}{\bar{x}^2 - \bar{x}^2} \right]}$$

$$q = F_{t(n-2)}^{-1} \left(1 - \frac{\alpha}{2} \right)$$

• Inferenza: intervallo di predizione per Y futuro

\tilde{x} valore di input di un prossimo punto (\tilde{x}, \tilde{Y}) $\tilde{Y} \sim \mathcal{N}(\beta_0 + \beta_1 \tilde{x}, \sigma^2)$
 → intervallo tubolare che contiene i punti (futuri)



Mentre l'intervalle di confidenza per la media si stringe sempre più, all'aumentare dei dati, quello di predizione per le singole risposte rimane abbastanza largo da racchiudere la maggior parte dei punti. La sua larghezza è circa $2 \cdot q \cdot \sigma$

→ vedi file `regr_lin_semplice.xlsx`

→ Calcoli per determinare l'intervallo di predizione

$$\tilde{Y} \approx \beta_0 + \beta_1 \tilde{x} \approx B_0 + B_1 \tilde{x} \quad \text{stima puntuale / centrale}$$

$$\Rightarrow \tilde{Y} - (B_0 + B_1 \tilde{x}) \sim \mathcal{N}(0; \sigma^2 + \frac{\sigma^2}{n} \left[1 + \frac{(\tilde{x} - \bar{x})^2}{\bar{x}^2 - \tilde{x}^2} \right])$$

↑
indip
↑
le var si sommano

$$\frac{\tilde{Y} - (B_0 + B_1 \tilde{x})}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(\tilde{x} - \bar{x})^2}{n(\bar{x}^2 - \tilde{x}^2)}}} \sim \mathcal{N}(0, 1) \Rightarrow$$

$$\frac{\tilde{Y} - (B_0 + B_1 \tilde{x})}{S_e \sqrt{1 + \frac{1}{n} + \frac{(\tilde{x} - \bar{x})^2}{n(\bar{x}^2 - \tilde{x}^2)}}} \sim t(n-2)$$

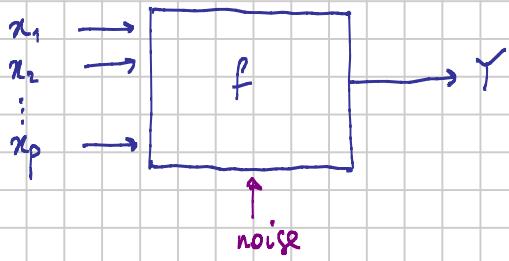
da qui si trova subito l'intervalle di predizione:

$$\tilde{Y} \in B_0 + B_1 x \pm q S_e \sqrt{1 + \frac{1}{n} \left[1 + \frac{(\tilde{x} - \bar{x})^2}{\bar{x}^2 - \tilde{x}^2} \right]} \quad q = F_{t(n-2)}^{-1} \left(1 - \frac{\alpha}{2} \right)$$

★ Esistono tante altre possibilità di fare inferenza

→ test e intervalli di confidenza su $\beta_0, \beta_1, \sigma, \beta_0 + \beta_1 x$

REGRESSIONE LINEARE Multipla



not. modello :

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + e$$

$e \sim N(0, \sigma^2)$

$$Y_i \sim N(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_p x_{i,p}; \sigma^2) \quad i = 1, 2, \dots, n$$

$$x_{i,0} = 1 \quad \text{"dummy" variable} \quad Y_i \sim N(\sum_{j=0}^p \beta_j x_{i,j}; \sigma^2)$$

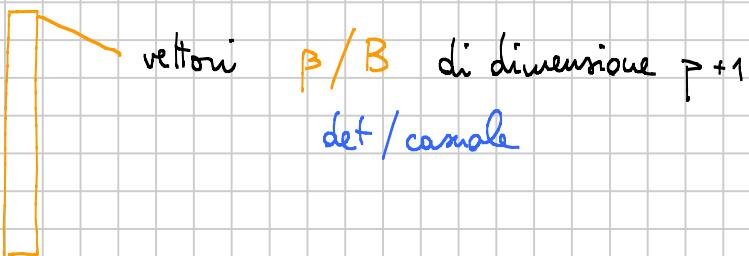
- In pratica: notazione matriciale

$$\sum_j x_{ij} \beta_j = [X\beta]_i$$

dataframe

id	x_0	x_1	x_2	\dots	x_i	\dots	x_p	Y
1	1	~	~					
2	1	~	~					
:	:							
i					x_{ij}			
:	:							
n	1							

vettore Y a valori in \mathbb{R}^n casuale



matrice $X \in M_{n,p+1}$ deterministica

- di solito $n \gg p$, perché se no è un problema
- modello

$$Y_i \sim N(\dots, \sigma^2) \quad E(Y_i) = \sum_{j=0}^p x_{ij} \beta_j = [X\beta]_i \quad \text{indipendenti al variare di } i$$

se considero Y un vettore aleatorio è un vettore Gaussiano con componenti indipendenti : $Y \sim N(X\beta; \sigma^2 I)$

- stimatori di $\beta \in \mathbb{R}^{p+1}$ e σ^2 : applica Cochran

$$E(Y) = X\beta = C\gamma$$

prendo $C = X$ e $\gamma = \beta$

$$\begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_{p+1} \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\hat{\beta} = B = (X^T X)^{-1} X^T Y$$



$$B \sim N(\beta; \sigma^2 (X^T X)^{-1})$$



• Stimatore di σ^2

$$SS_R = \|Y - XB\|^2 = \sum_{i=1}^n \left(Y_i - \sum_{j=0}^p X_{ij} B_j \right)^2$$

$$S_e := \sqrt{\frac{SS_R}{n-p-1}}$$

$$\frac{S_e^2}{\sigma^2} (n-p-1) = \frac{SS_R}{\sigma^2} \sim \chi^2(n-p-1)$$

indip da B

HW: Applicare Cochran a questo caso:

$$Y_1, \dots, Y_{n_1}, Y_{n_1+1}, \dots, Y_{n_1+n_2}, \dots, Y_{n_1+n_2+\dots+n_k}$$

↓ k campioni ↑ in fila
 $N(\mu_1, \sigma^2)$ $N(\mu_2, \sigma^2)$ $N(\mu_k, \sigma^2)$

• Test per il potere predittivo delle singole variabili di ingresso

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + e$$

Per $j = 1, 2, \dots, p$ voglio verificare se Y dipende davvero da x_j

$$H_0: \beta_j = 0 \quad H_1: \beta_j \neq 0 \quad \text{ipotesi del test}$$

Parametro: $\beta_j \rightarrow$ stimazione: $B_j \rightarrow$ distribuzione: $B_j \sim ?$

$$\text{So che } B \sim N(\beta; \sigma^2(X^T X)^{-1}) \Rightarrow B_j \sim N(\beta_j; \sigma^2 [(X^T X)^{-1}]_{jj})$$

$$\rightarrow \text{funz. anc: } \frac{B_j - \beta_j}{\sigma \sqrt{[(X^T X)^{-1}]_{jj}}} \sim N(0, 1)$$

le varianze sono sulla diagonale

$$\frac{B_j - \beta_j}{\sigma \sqrt{[(X^T X)^{-1}]_{jj}}} \sim t(n-p-1) \quad \text{funz. anc.} \rightarrow \text{stat.: } T_j := \frac{B_j}{S_e \sqrt{[(X^T X)^{-1}]_{jj}}} \stackrel{H_0}{\sim} t(n-p-1)$$

perche' Se indip. da B

$$T_j := \frac{B_j}{S_e \sqrt{[(X^T X)^{-1}]_{jj}}} \stackrel{H_0}{\sim} t(n-p-1)$$

* La statistica T_j si chiama anche coefficiente normalizzato

- sotto H_0 ha legge $t(n-p-1)$ (valori vicini a 0)
- sotto H_1 ha legge di media $\neq 0$

* Il test si può fare calcolando:

$$- \text{la R.A.} = [-q \downarrow i + q] \quad q = F_{t(n-p-1)}^{-1} \left(1 - \frac{\alpha}{2} \right)$$

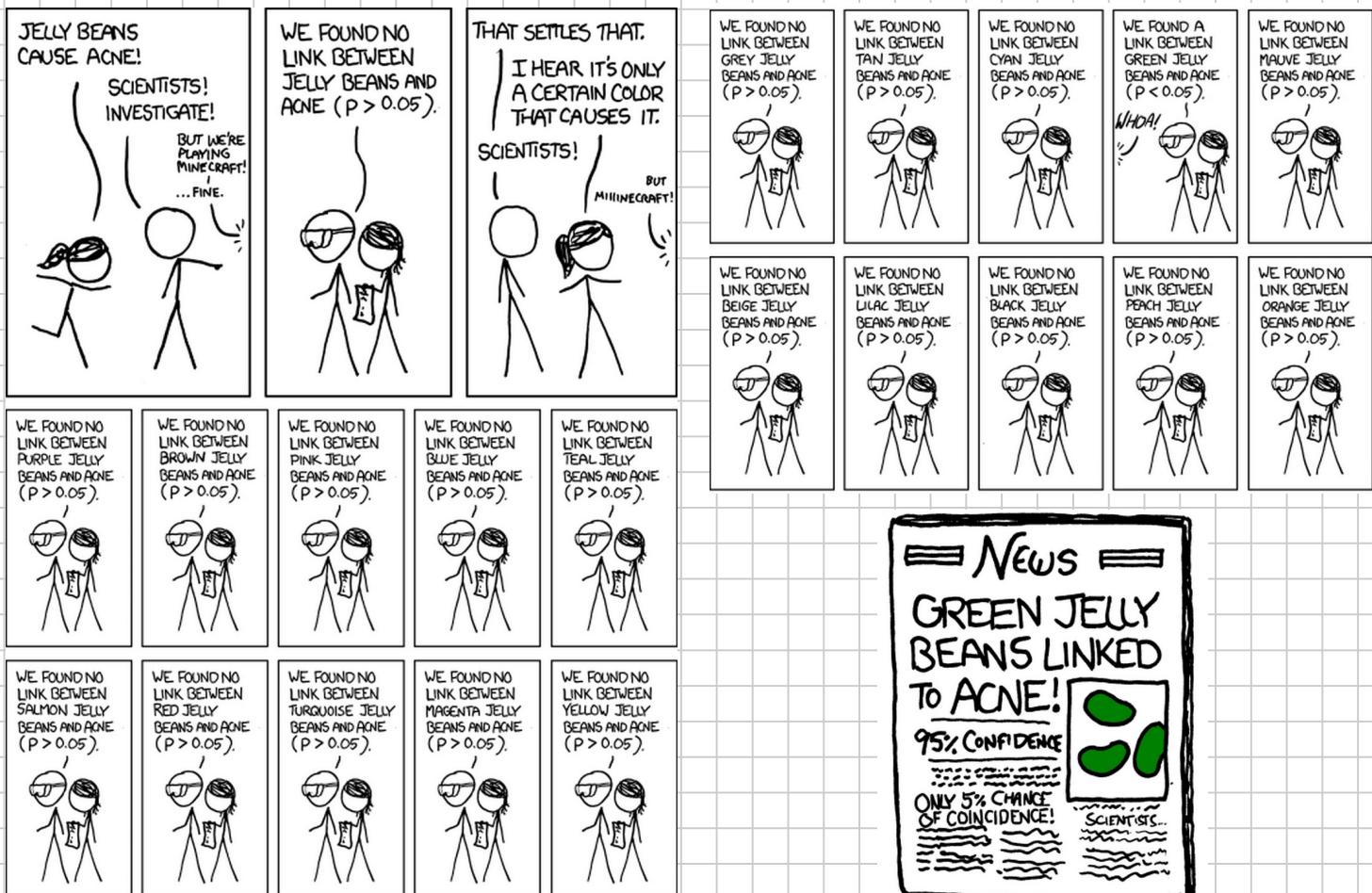
$$- \text{il p-value} \quad x_j^* = 2 - 2 F_{t(n-p-1)}(|T_j|)$$

bil di significatività

* Come concludo? Dipende da diverse cose...

- $x_j^* \geq 30\%$ molto grande: x_j ha potere predittivo trascurabile \rightarrow pensa di rimuoverla
- $x_j^* \leq 0.1\%$ molto piccolo: x_j ha chiaro potere predittivo
- $0.1\% < x_j^* < 30\%$ dipende dal task, dalla situazione, dal tipo di selezione

PROBLEMA DEI TEST MULTIPLI



α : livello di significatività, tipicamente 5%
upper bound sulla probabilità di errore di I specie α

→ se faccio un solo test va bene

→ se faccio molti test è probabile che almeno uno di essi compatti un errore di I specie

$$\begin{aligned} & P(\text{almeno errore di I specie} \mid \text{tutti i test vera } H_0) \\ &= P\left(\bigcup_{i=1}^n \{\text{dico } H_1 \text{ nel test } i\} \mid \text{tutti i test vera } H_0\right) \\ &\leq \sum_{i=1}^n P(\text{errore I specie nel test } i \mid H_0 \text{ vera nel test } i) \end{aligned}$$

Se $\bar{\alpha}_i = \alpha$ $i=1, 2, \dots, n$ allora $\alpha \text{ globale} \leq n \alpha$

• CORREZIONE DI BONFERRONI : fisso $\bar{\alpha}$ globale (intorno al 5%)

e poi pongo $\bar{\alpha}_i = \frac{1}{n} \bar{\alpha}$ $i=1, 2, \dots, n$

→ questo garantisce che ci sia una probabilità di errori di I specie (falsi positivi) controllata

→ ha il prezzo altissimo di abbassare la potenza dei test

* In generale è bene prima di fare i test ridurne il più possibile il numero, per diminuire questo problema.

* Assolutamente decidere i test anche prima di vedere i dati (ed evitare il "cherry-picking")

■ SELEZIONE DELLE VARIABILI

■ METODI STEPWISE

- backward : si parte da p variabili e si scende di 1 alla volta
- forward : si parte da 0 variabili e si sale di 1 alla volta
- misto : come forward, ma è possibile inserire passi backward

- Backward : si guardano i p -value α_j^* delle variabili rimanenti
(il test è $H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$)

→ se $\alpha_j^* < 0.001$ sono sicuri che $\beta_j \neq 0$

(non la tolgo di certo) → correzione di Bonferroni $\approx \frac{5\%}{p}$

→ se $0.1\% \leq \alpha_j^* < 30\%$ non so (di solito le tengo)

→ se $\alpha_j^* \geq 30\%$ non vi è alcun giudizio che $\beta_j \neq 0$
posso pensare di togliere la variabile

→ in pratica, la soglia suggerita tipicamente è 30%

→ se tutte le variabili rimanenti sono significative ($\alpha^* <$ soglia)
interrumpo la procedura

→ altrimenti tengo una variabile fra quelle con $\alpha^* >$ soglia

↳ di solito quella con α^* minimo è la scelta naturale

↳ futuraria sarebbe bene provare con diverse variabili e confrontare un po' i risultati per scegliere quella da togliere

* Non è detto che la variabile con α^* più alto sia la meno utile

↳ infatti quando è vera H_0 , $\alpha^* \sim \text{unif}(0,1)$, non predilige per forza i valori vicini a 1.

→ selezione stepwise backward risente di più delle multicollinearità

	pizzone	x_1 : fatturato	x_2 : # dipendenti	x_3 : importi	x_4 : # posti a sedere	Y : arricchimento
1		140	6	120	30	7
2		100	4	90	25	10
...						
n		260	11	230	50	9

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \dots + \epsilon$$

$$x_1 \approx 22 x_2 \quad x_3 \approx 20 x_2 \quad x_4 \approx 5 x_2$$

$$Y = 3 + 0.1 x_1 + 1 \cdot x_2 + 0.1 x_3 + \dots$$

$$\approx 3 + (0.1 \cdot 22 + 1 + 0.1 \cdot 20) x_2 + \dots$$

$$= 3 + (0 \cdot 22 + 3.2 + 0.1 \cdot 20) x_2 + \dots$$

$$\approx 3 + 0 x_1 + 3.2 x_2 + 0.1 x_3 + \dots$$

→ i coefficienti $\beta_1 - \beta_4$ sono molto imprecisi

↳ $[(X^T X)^{-1}]_{jj}$ sarà grande

↳ T_j sarà piccolo

↳ κ_j^* sarà grande : le var $x_1 - x_4$ avranno $\kappa^* >> 30\%$

→ se seguendo la procedura "naturale" seleziona una delle quattro a caso

↳ se sono presenti forti correlazioni fra variabili di ingresso,

i corrispondenti stimatori β_j hanno varianza elevata e spesso

i p-value risultano alti fino a che si riduce la correlazione

togliendo delle variabili.

→ se il range di X non è max, $X^T X$ non è invertibile; se c'è correlazione, è instabile

→ inoltre nel backward all'inizio il numero di variabili è massimo = p

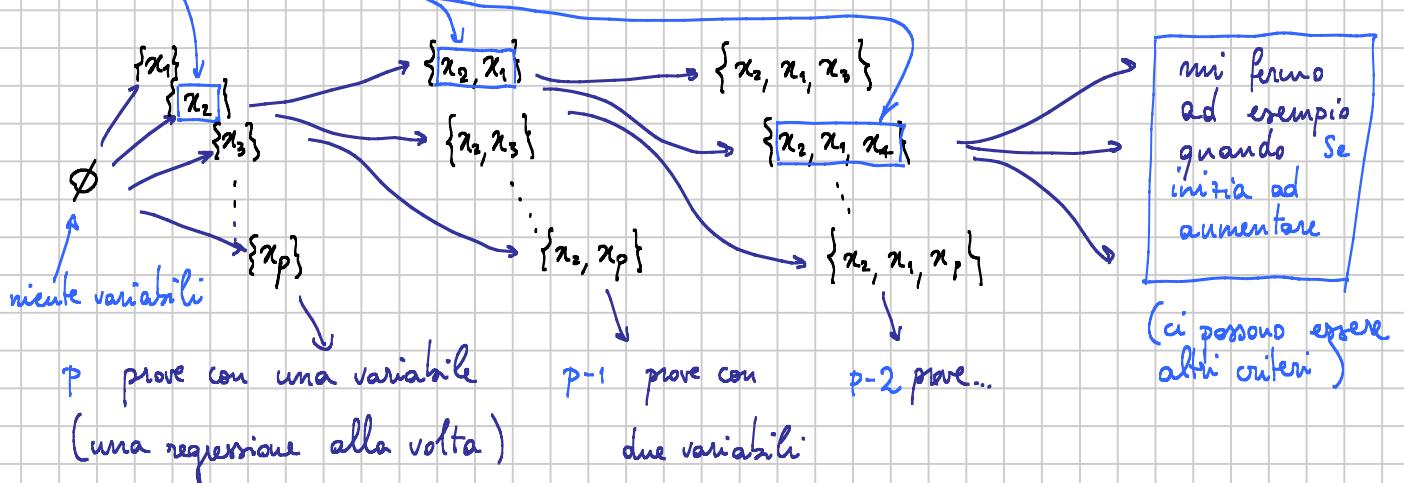
↳ $\frac{n}{p} \epsilon$ minimo → regressione meno robusta

- Forward : Si confrontano i modelli ottenuti aggiungendo una sola delle variabili non ancora incluse

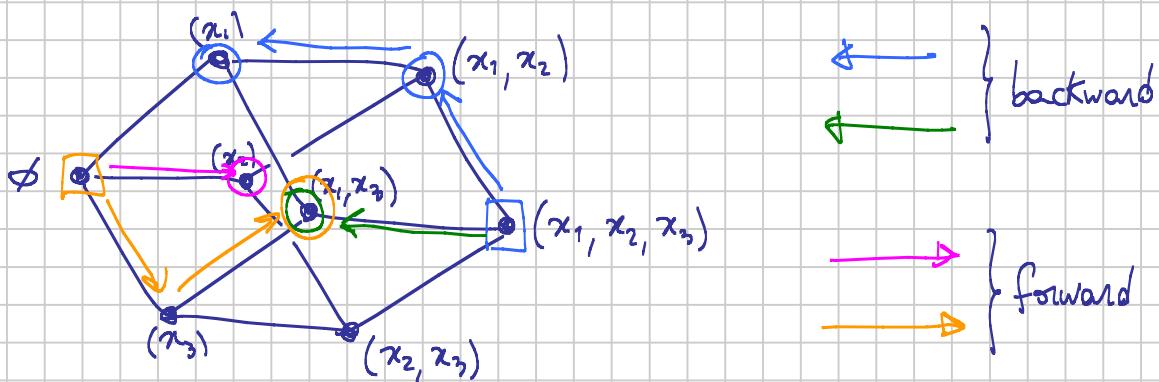
si procede per tentativi, guidati da indicatori globali come Se

I step : scelgo quelle con α_j^* minimo o Se minimo o R^2 massimo ... è equivalente

step successivi : scelgo quella che risulta in Se minimo o R^2 massimo : non posso più guardare α_j^* perché sono due o più



- Forward e backward a volte si incontrano, a volte no



- È comune che i software che automatizzano la stepwise chiedano di fissare una soglia per il valore di F invece che per α^*

- "F to enter" : forward } i valori di default di solito
- "F to remove" : backward } corrispondono a $\alpha^* \geq 30\%$
- entrambi nella stepwise mista

METODI GLOBALI

prevedono di testare tutti i sottosinsiemi di variabili (sono 2^p , è fattibile per p non troppo grande)

→ per ciascuno provo la regressione

→ basta avere uno (o più) "score" per confrontare i modelli

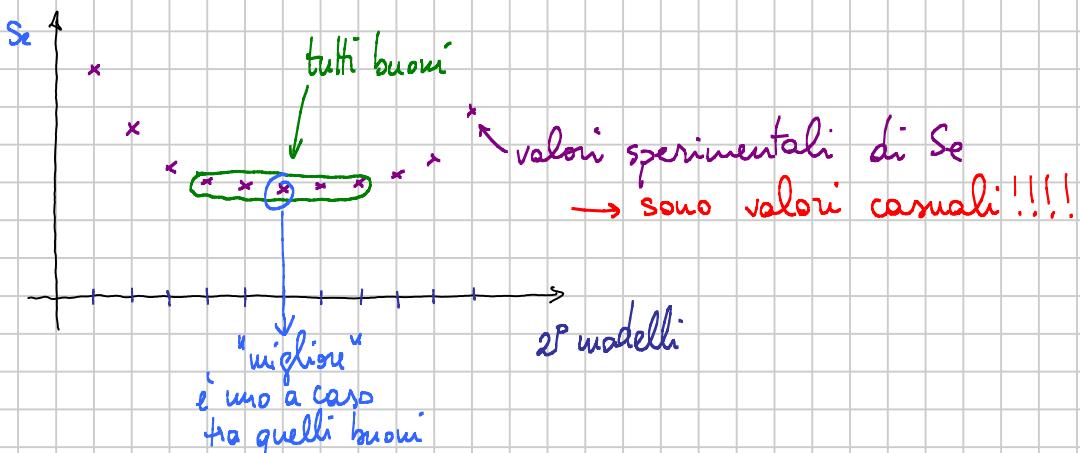
* Quali score?

a) $Se \approx \sigma$ minore è, meglio è

a') R_A^2 coeff. di determinazione corretto maggiore è, meglio è (equiv. a Se)

b) AIC = $2k - 2\log(\text{likelihood})$ minore è, meglio è

c) Validazione (vedi oltre) su Se , SS_R , R_D^2 , ...



* Non è opportuno cercare il minimo di una funzione usando una sua stima casuale

• Molte volte tutti i metodi coincidono

ora 23

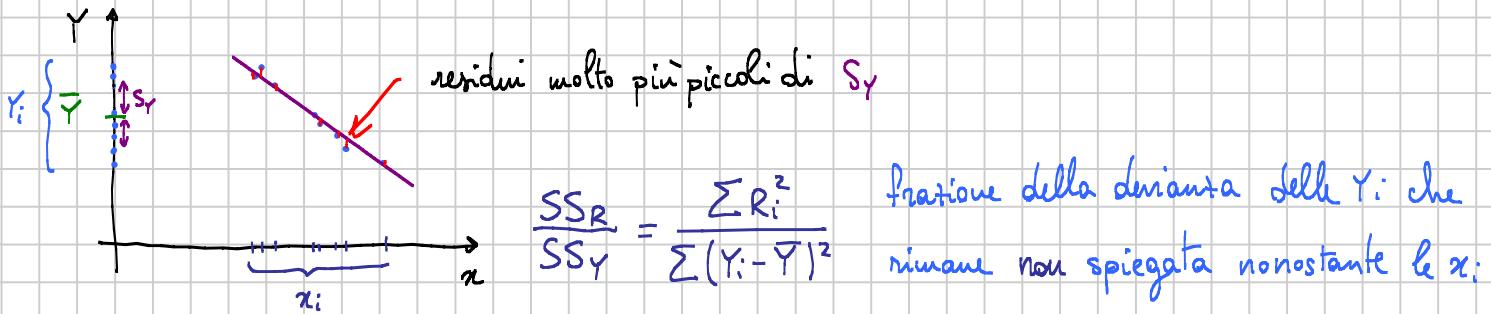
COEFFICIENTI DI DETERMINAZIONE R_D^2 E R_A^2

* Coefficiente di determinazione $R_D^2 \in [0, 1]$

$$R_D^2 := 1 - \frac{SS_R}{SS_Y}$$

frazione della devianza spiegata dal modello

dove $SS_Y := \sum_{i=1}^n (Y_i - \bar{Y})^2$, $SS_R := \sum_{i=1}^n (Y_i - [\bar{X}B]_i)^2 = \sum_{i=1}^n R_i^2$



$$R_D^2 \text{ max} \Leftrightarrow \|R\|^2 \text{ min}$$

- al crescere di p , $\|R\|^2$ diminuisce sempre, R_D^2 aumenta sempre
- $\hookrightarrow R_D^2 \text{ max} \Leftrightarrow$ tutte le vars
- \hookrightarrow non va bene per la selezione

$$SSR(\{1, 2, \dots, p\}) = \min_{(B_0, \dots, B_p)} \sum_i (Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \dots - B_p x_{ip})^2$$

\nwarrow

$$SSR(\{1, 2, \dots, p-1\}) = \min_{(B_0, \dots, B_{p-1})} \sum_i (Y_i - B_0 - B_1 x_{i1} - \dots - B_{p-1} x_{ip-1} - 0 x_{ip})^2$$

\uparrow minimo con $B_p = 0$

* Coefficiente di determinazione corretto $R_A^2 \in [0, 1]$ (adjusted)

$$R_A^2 = 1 - \frac{S_e^2}{S_Y^2}$$

dove $S_e^2 := \frac{SSR}{n-p-1}$ e $S_Y^2 := \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{1}{n-1} SSY$

$$R_A^2 \text{ max} \Leftrightarrow S_e^2 \text{ min} \quad (\text{non è monotono con le vars})$$

* Si ha sempre $0 \leq R_A^2 \leq R_D^2 \leq 1$ e $R_A^2 = 1 \Leftrightarrow R_D^2 = 1$

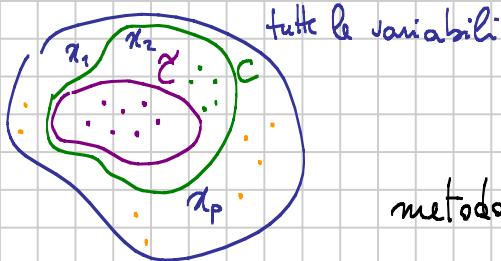
HW: Trovare $a=a(n,p)$, $b=b(n,p)$: $R_A^2 = a + b R_D^2$
e mostrare i due fatti qui sopra

* Quando $p=1$ (regr. lin. semplice) si trova che R_D^2 è legato al coefficiente di correlazione lineare fra x e Y

$$Scorr(x, Y)^2 = R_D^2$$

HW: check

• Un primo esempio di ANOVA (analisi della varianza)



$$\bar{C} \subseteq C \subseteq \{x_1, x_2, \dots, x_p\} \quad \#\bar{C} = d \quad \#C = d$$

candidati insiemi di variabili

metodo per confrontare i modelli che usano C o \bar{C}

→ faccio le due regressioni $R_D^2(C) \geq R_D^2(\bar{C}) \Leftrightarrow SS_R(C) \leq SS_R(\bar{C})$

$$SS_D := SS_R(\bar{C}) - SS_R(C)$$

\uparrow
 $d-d$

\uparrow
 $n-\bar{d}-1$

\uparrow
 $n-d-1$

$$\frac{1}{\sigma^2} SS_R(\{x_1, \dots, x_p\}) \sim \chi^2(n-p-1)$$

$$\frac{1}{\sigma^2} SS_R(\bar{C}) \sim \chi^2(n-\bar{d}-1)$$

$$\frac{1}{\sigma^2} SS_R(C) \sim \chi^2(n-d-1)$$

$$\chi^2(n-d-1) + \chi^2(d-\bar{d}) = \chi^2(n-\bar{d}-1)$$

\uparrow
 indip

→ Calcolo la statistica :

$$V := \frac{SS_D / (d-\bar{d})}{SS_R(C) / (n-d-1)}$$

→ Thm (si dimostra con Cochran) :

sotto l'ipotesi nulla che le variabili in più non siano utili,

$$H_0: \beta_j = 0 \quad \forall j \in C \setminus \bar{C}$$

$$\frac{1}{\sigma^2} SS_D \sim \chi^2(d-\bar{d}) \text{ indip da } SS_R(C)$$

la statistica ha legge F di Fisher :

$$V \stackrel{H_0}{\sim} F(d-\bar{d}; n-d-1)$$

→ Invece sotto l'ipotesi alternativa, V è tipicamente grande,
perciò il test è unilaterale $RA_V = [0; b]$ $b = F_{F(.,.)}^{-1}(1-\alpha)$

$$\alpha^* = 1 - F_{F(.,.)}(V)$$

HW : cosa succede se $\bar{C} = \emptyset$ e $C = \{x_1, \dots, x_p\}$?

(test globale di regressione : di solito α^* molto molto piccolo)

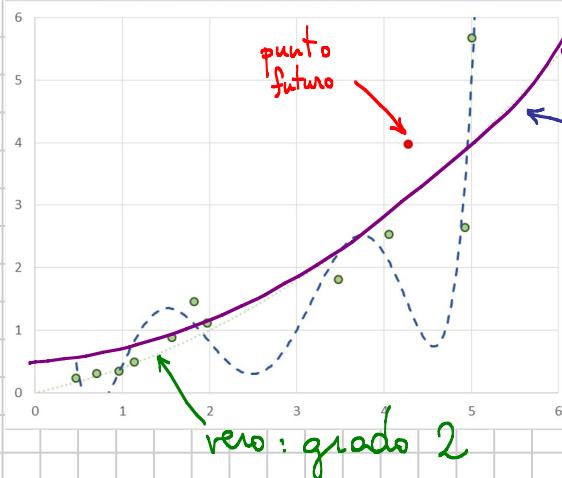
* Altro caso : $C = \bar{C} \cup \{x_j\}$ (stepwise forward o backward)

test alternativo a quello canonico $H_0: \beta_j = 0 \quad H_1: \beta_j \neq 0$

che usa la F di Fisher invece della t di Student (ma è equivalente)

OVERFITTING

- è un problema universale del machine learning, dalla regressione alle reti neurali
- si ha quando il modello ha troppe variabili (o coefficienti incogniti) rispetto al numero di punti



$$Y = \tilde{B}_0 + \tilde{B}_1 x + \tilde{B}_2 x^2$$

$$Y = B_0 + B_1 x + B_2 x^2 + \dots + B_p x^p$$

la curva di grado elevato approssima meglio dell'altra i punti del dataset, ma peggio (molto peggio) i punti futuri: non generalizza

→ Nella regressione è meglio se $\frac{N}{P} \gg 1$ per evitare overfitting

→ $(X^T X)^{-1}$ che è la covarianza dei B_j tende ad allontanarsi da una matrice diagonale → gli stimatori B_{j*} sono sempre più correlati, meno precisi → i test sulla significatività di B_{j*} sono meno potenti → molti x_{j*} sono alti

→ per questo è meglio la forward, che incrementando $j = 1, 2, \dots$ evita l'overfitting.

→ per evitare la correlazione fra le variabili che rende non invertibile e non univoca la selezione, va progettato l'esperimento fissando in anticipo i valori di input

DoE Design of Experiment (Sleicer "60 stat...", cap 10)

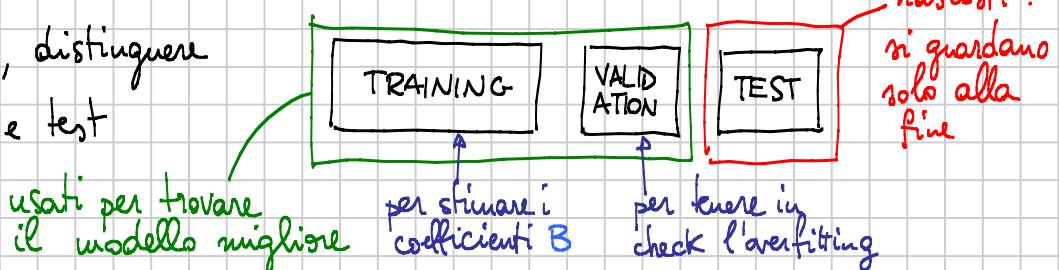
esempio di design
"ortogonale"
con tre variabili

x_1	x_2	x_3
-1	-1	-1
-1	+1	+1
+1	-1	+1
+1	+1	-1

-1 e +1 sono due qualsunque valori fissati, diversi per le tre variabili

Validazione

- serve ad evitare l'overfitting
 - metto da parte una frazione casuale dei dati (10-20%, validation set)
 - faccio la regressione sugli altri (train set)
 - verifico che i residui di regressione non siano molto minori di quelli che ha il modello con i dati di validazione
 - se non è così sono in overfitting: devo ridurre le variabili e riprovare
 - i dati di validazione permettono di misurare le performance del modello in modo onesto e senza bias, e possono guidare la scelta di variabili (selezione) e iperparametri (vedi offre).
 - il coefficiente di determinazione calcolato sui dati di validazione non è monotono con il numero di variabili
 - se possibile, distinguere validazione e test



• Varianti:

→ cross-validation: $\frac{1}{2}$ train $\frac{1}{2}$ validation e poi scambio

voglio calcolare R^2 sul set di validazione invece che su quello di training → vantaggio: così misura davvero quanto è buono il modello

$$R_D^2 = 1 - \frac{SS_R^{CV}}{SS_Y}$$

$$SS_R^{CV} = \sum_{i \in A_2} (Y_i - [XB^{A_1}]_i)^2 + \sum_{i \in A_1} (Y_i - [XB^{A_2}]_i)^2$$

→ k-fold cross-validation: diviso in k blocchi, ne uso k-1 di training e 1 di validazione e ciclo

↳ per k=2 viene la cross-validation

→ jackknife: k-fold con k=n

Regolarizzazione

- spesso in caso di overfitting la funzione approssimatrice ha derivate molto grandi \rightarrow coefficienti molto grandi: l'idea è di penalizzarli

- regressione standard:

LSE minimizzo $SS_R = \sum_{i=1}^n (Y_i - \sum_{j=0}^P B_j x_{ij})^2$

- ridge regression:

minimizzo $\frac{1}{n} \sum_{i=1}^n (Y_i - \sum_{j=0}^P B_j x_{ij})^2 + \alpha \cdot \sum_{j=0}^P \frac{B_j^2}{S_{xj}^2}$

- lasso regression:

minimizzo $\frac{1}{n} \sum_{i=1}^n (Y_i - \sum_{j=0}^P B_j x_{ij})^2 + \lambda \cdot \sum_{j=0}^P \frac{|B_j|}{S_{xj}}$

\rightarrow con questo alcuni B_j vanno a 0: ottengo una relazione variabili

- elastic regression: ci sono entrambe le penalizzazioni

* In genere questi modelli si risolvono con un ottimizzatore iterativo e non in modo analitico.

* Gli iperparametri λ , α vanno scelti con cura: di solito si usa il set di validazione e si fanno diverse prove.

* E' fondamentale, perché questi termini abbiano senso, che i coefficienti B_j abbiano scala confrontabile: di solito ci si accontenta di standardizzare preliminarmente le variabili (vedi oltre).

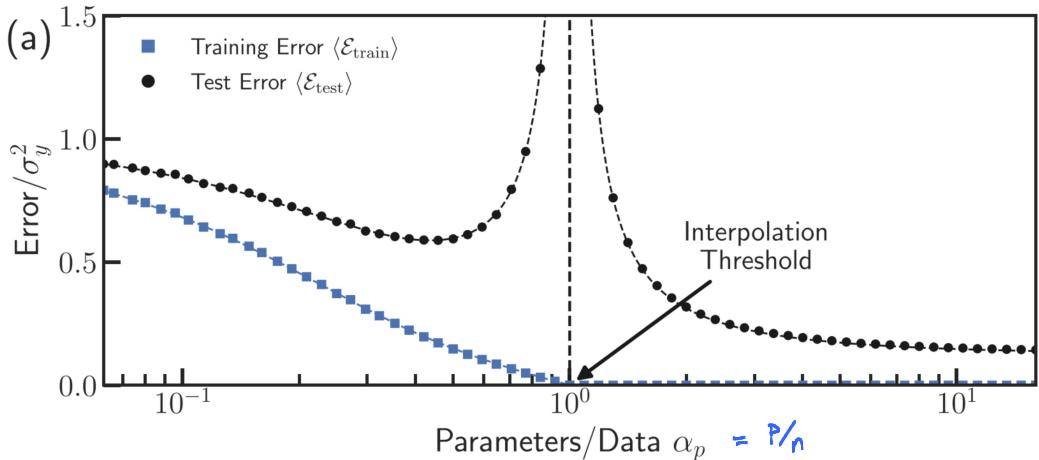
\rightarrow LASSO regression (least absolute shrinkage and selection operator)

minimizzo SS_R con il vincolo: $\sum_{j=1}^P |B_j| \leq \alpha$

\rightarrow RIDGE regression: tipo LASSO con $\alpha = 2$

■ DOUBLE DESCENT

In ambiti di machine learning realistici (ottimizzazione iterativa + regolarizzazione implicita), quando il rapporto n/p è minore di 1 e tende a 0 l'overfitting scende!



● Regressione polinomiale

se il modello lineare non mi soddisfa ...

(tipicamente perché i residui di regressione mostrano andamenti nonlineari rispetto a qualcuna delle variabili)

... posso aggiungere termini nonlineari : dummy che sono monomi di grado 2 o più

Ad es : se $p=1$

$$\text{lineare} : Y = \beta_0 + \beta_1 x + e$$

$$\text{quadratico} : Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

è ancora lineare in β

$$\text{generale} : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \dots + \beta_d x_d^d$$

x_1 x_2 x_d

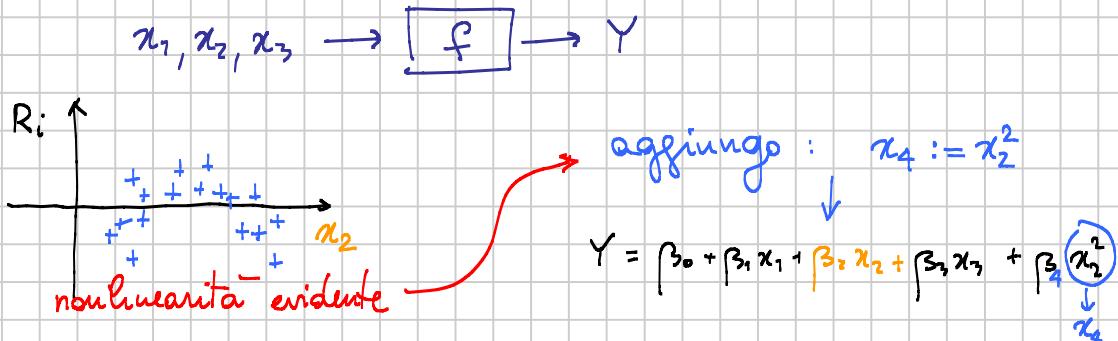
$$C = \begin{pmatrix} 1 & x(1) & x(1)^2 \\ 1 & x(2) & x(2)^2 \\ \vdots & \vdots & \vdots \\ 1 & x(n) & x(n)^2 \end{pmatrix}$$

facciamo finita riassumendo così

id	x_0	x_1	x_1^2	\dots	x_d
1	1	$\xrightarrow{\sim}$	\sim	\sim	
:	:	$\xrightarrow{\sim}$	\sim	\sim	
:	:	$\xrightarrow{\sim}$	\sim	\sim	
n	1	$\xrightarrow{\sim}$	\sim	\sim	

è esattamente una regressione lineare multipla, solo con dummy variables, invece che dati reali

- * L'interpolazione polinomiale funziona così
- * Regola gerarchica: non si tolgono gradi intermedi, si guarda solo il meggiore
- * Attenzione che è meglio forward che backward (solo variabili correlate)

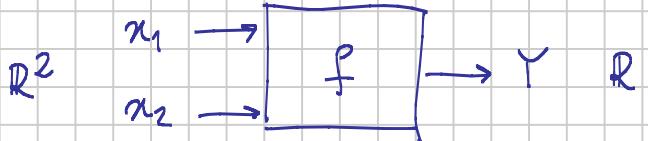


la nonlinearietà si corregge, x_4 risulterà significativa, SS_R sarà minore

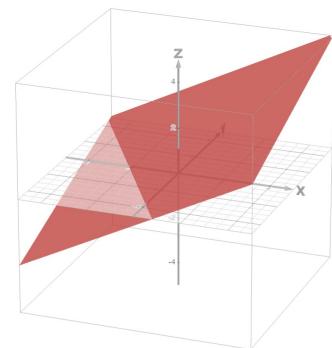
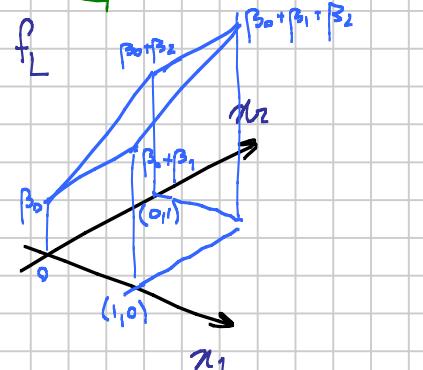
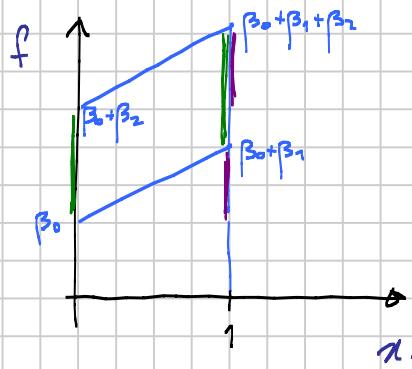
→ Altri esempi: $[x_1 x_2]$, termini di interazione $x_2^3, x_1 x_2^2, \dots$

→ Regola gerarchica: si tengono tutte le variabili corrispondenti
a monomi che dicono un monomio del modello: $x_1 x_2 \rightarrow x_1 x_2, x_1^2, x_2^2$

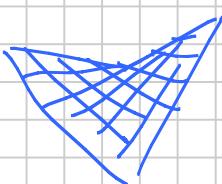
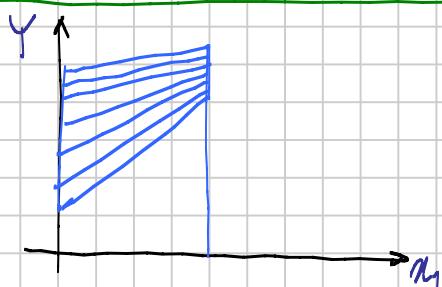
• Termini di interazione



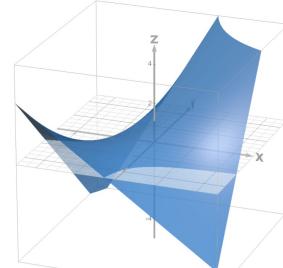
modello lineare (senza $x_1 x_2$)



modello semilineare (con $x_1 x_2$)



$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$



→ Il modo in cui Y dipende da x_1 , dipende a sua volta da x_2 (e viceversa)

• Selezione delle variabili in modelli nonlineari multipli

- backward : i. modello lineare con tutte le var
 ii. controllo i residui per non-linearità evidenti } provo ad aggiungere
 iii. controllo se ci sono vars con p-value infino } termini nonlineari
 iv. comincio a togliere variabili, tenendo conto della regola
 gerarchica, e controllando sempre i residui

- forward : i. scelgo la prima variabile lineare $\rightarrow Y = \beta_0 + \beta_1 x_1$
 ii. scelgo la seconda fra x_2, x_3, \dots, x_p e x_1^2
 iii. prosegua così
 \hookrightarrow se ho inserito x_1, x_2 , cercherò fra $x_3 - x_p, x_1^2, x_2^2, x_1 x_2$

■ REGRESSIONE PESATA

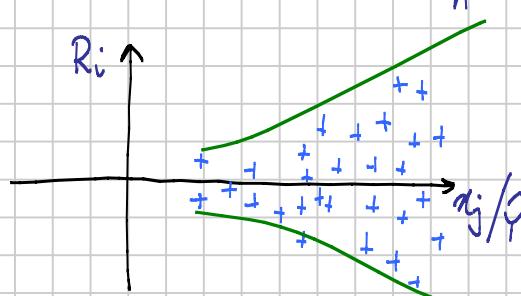
si usa quando i dati non sono omoschedastici

→ se $Y_i \sim \text{Pois}(\mu_i) \sim \mathcal{N}(\mu_i; \mu_i)$ non lo sono $\rightarrow Y_i \sim \mathcal{N}(\sum \beta_j x_{ij}; \sigma^2)$
 \hookrightarrow varianza $\propto \mu \approx \hat{\mu}$

→ se $Y_i \sim \mathcal{N}(\mu_i; ?)$ con varianza tale che $\hat{Y}_i = \mu_i \pm \underbrace{10\% \mu_i}_{\text{dev. std.}}$ non lo sono
 \hookrightarrow dev. std. $\propto \mu \Rightarrow$ varianza $\propto \mu^2 \approx \hat{\mu}^2$

→ se $Y_i \sim \text{bin}(n_i; p_i) \Rightarrow \tilde{Y}_i := \frac{1}{n_i} Y_i \sim \mathcal{N}(p_i; \frac{p_i(1-p_i)}{n_i})$ non lo sono
 \hookrightarrow varianza $\propto \frac{\mu(1-\mu)}{n} \approx \frac{\hat{\mu}(1-\hat{\mu})}{n}$

→ se i residui ...



$$\begin{aligned}\sigma &\propto x_j \\ \sigma^2 &\propto x_j \\ \sigma &\propto \hat{\mu} \\ \sigma^2 &\propto \hat{\mu}\end{aligned}$$

Si può risolvere il modello se suppongo $\boxed{\sigma_i \propto r_i}$, noti

$$\text{Likelihood} : L(\beta) = \prod_{i=1}^n f(y_i; x_i, \beta)$$

$$l(\beta) = \log L(\beta) = C - \sum_{i=1}^n \frac{(y_i - \sum_{j=0}^p x_{ij} \beta_j)^2}{2 \sigma_i^2}$$

- caso $\sigma_i = \sigma$ $l(\beta)$ max se $\sum_{i=1}^n (Y_i - \sum_j x_{ij} B_j)^2 =: SS_R$ min

- caso σ_i generico $l(\beta)$ max se $\sum_{i=1}^n w_i (Y_i - \sum_{j=0}^p x_{ij} B_j)^2 =: \tilde{SS}_R$ min

$$w_i \propto \frac{1}{\sigma_i^2} \propto \frac{1}{r_i^2} \quad w_i = \frac{1}{r_i^2}$$

→ Metodo equivalente:

risolvere la regressione con \tilde{X} e \tilde{Y} modificati così:

$$\tilde{x}_{ij} := \frac{x_{ij}}{r_i} = \sqrt{w_i} x_{ij} \quad i = 1, 2, \dots, n \quad j = 0, 1, 2, \dots, p \quad \text{va incluso}$$

$$\tilde{y}_i := \frac{y_i}{r_i} = \sqrt{w_i} Y_i \quad X_{i,0} = 1 \quad \boxed{\tilde{x}_{i,0} = \frac{1}{r_i} = \sqrt{w_i}}$$

$$\tilde{SS}_R := \sum_{i=1}^n \left(\tilde{y}_i - \sum_{j=0}^p \tilde{x}_{ij} B_j \right)^2 = \sum_{i=1}^n \frac{1}{r_i^2} \left(Y_i - \sum_{j=0}^p x_{ij} B_j \right)^2 = \sum_{i=1}^n w_i (Y_i - \sum_{j=0}^p x_{ij} B_j)^2$$

INTRODUZIONE AL MACHINE LEARNING

Note Title

ora 25

29/04/2025

GESTIONE VARIABILI

(regressione ma non solo)

- Tipologie: dicotomiche, categoriche, numeriche

Dicotomiche

come variabili di ingresso, codificate con qualunque coppia di numeri

→ tipicamente userò 0 e 1

feum	x_j
1	
0	
1	
1	
0	
:	

$$Y = \sum_{j=0}^p \beta_j x_j + e$$

β_j / B_j : differenza di risposta (Y) che hanno le feum rispetto ai mas

→ sono le variabili scelte nel D.o.E.!

come variabile di risposta (nella regressione) esistono:

↳ discriminant analysis (regressione standard fissa ipotesi)

↳ regressione logistica (regressione con MLE Bernoulliano vedere oltre)

• categoriche : nominali (20 regioni, stagioni, tipo di contaminante, nome, ...)
CIFAR-10 { cane, gatto, camion, ... } , { '0', '1', '2', ..., '9' })

ributta

name

ordinali (titolo di studio, classifica (I, II, III), gravità di sintomi, ...)

* Le ordinali posso decidere di codificarle come numeri, rispettando l'ordine e sperando che funzion... .

titolo di studio :	niente	1	0	0	anni di studio
	elementari	2	10	5	
	medie	3	20	8	
	maturità	4	= 30	13	
	laurea	5	40	16	
	laurea magistrale	6	50	18	
	dottorato	7	60	21	
	front line non cambia				un po' diverse

* Negli altri casi :

[difficile] come variabili di risposta

(ma esistono regr. logistica, classification trees, random forests e reti neurali)

Come variabili di ingresso :

🚩 **[ok]** con la regressione, ma le categorie (**k**) vanno esplose in **k-1** dicotomiche (se non sono linearmente dipendenti)
→ codifica one-hot con una categoria tralasciata

🚩 **[ok]** se uso l'ANOVA (tipi regr. lin. semplice, ma con χ categorica)
* se le variabili di ingresso sono:
a) una sola ; categorica → ANOVA a 1 via
b) due sole ; categoriche → ANOVA a 2 vie

• Codifica one-hot di variabili di ingresso categoriche nominali
→ cose da sapere

x_0		pri	est	aut	inv
1	primavera	1	0	0	0
1	estate	0	1	0	0
1	autunno	0	0	1	0
1	i nverno	0	0	0	1

default

*codifica univoca
"one-hot"*

→ equazione

$$Y = \beta_0 + \beta_p x_p + \beta_e x_e + \beta_a x_a + \dots + e$$

$$Y = \beta_0 + \beta_p + \dots \quad \text{primavera}$$

$$Y = \beta_0 + \beta_e + \dots \quad \text{estate}$$

$$Y = \beta_0 + \beta_a + \dots \quad \text{autunno}$$

$$Y = \beta_0 + \dots \quad \text{i*nverno*}$$

β_a è diff. aut-inv

→ i coefficienti misurano l'effetto delle categorie rispetto al default

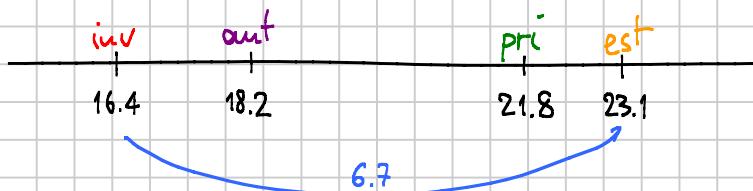
• Come si interpretano i coefficienti B corrispondenti

→ esempio : quattro stagioni, tre coefficienti

$$Y = B_0 + B_1 x_1 + B_2 x_2 + B_3 x_3 + \dots$$

↑ ↑ ↑ ↑
consumi pri est aut

$$B_0 = 16.4 \quad B_1 = 5.4 \quad B_2 = 6.7 \quad B_3 = 1.8$$



$$H_0: \beta_3 = 0$$

vuo dire che

autunno \approx inverno

$$H_0: \beta_1 = \beta_2 \quad \text{vuo dire che primavera} \approx \text{estate}$$

* Se fa stepwise backward elimina una di queste dicotomiche
vuo dire che non la distingue da quella di default.

L'effetto è automaticamente che le due categorie si fondono

→ se autunno è "non significativa", ovvero il test dice $H_0: \beta_3 = 0$

	PRI	EST
primavera	1	0
estate	0	1
autunno	0	0
inverno	0	0

categoria autunno - inverno

→ se voglio confrontare primavera con estate, ho due scelte :

a. faccio un test manuale $H_0: \beta_1 = \beta_2 \quad H_1: \beta_1 \neq \beta_2$

HW: scrivere il test (statistica, p-value)

b. cambio il default (metto estate come default) e rifaccio

* Attenzione! Sale (molto) p e peggiora (a volte troppo) il rapporto n/p

→ per ridurre il problema, si possono accoppare categorie troppo fici in macro categorie, o anche buttare un po' di righe del dataset

* Nonlinearità : $0^2 = 0 \quad 1^2 = 1$ non serve a niente aggiungere potenze
 $x_{\text{rest}} \cdot x_{\text{pri}} = 0$ inutile
 $x_{\text{rest}} \cdot x_i \rightarrow \text{ok}$ non della stessa categoria

* Stepwise forward : sceglie lui i / il default e non è detto che lo faccia nel modo migliore

ad es: $\emptyset \rightarrow x_{\text{rest}} \rightarrow x_{\text{rest}}, x_{\text{pri}}$ stop

magari più - est possono essere collassate e lui non può accorgersene

• Numeriche : tipo "differente" (non faccio niente)

(Q.I., peso, statura, consumi, ...)

tipo "rapporto" (faccio il logaritmo e poi vado avanti)

sono positive, coprono vari ordini di grandezza

(reddito, abitanti città, intensità sonora (dB), ...)

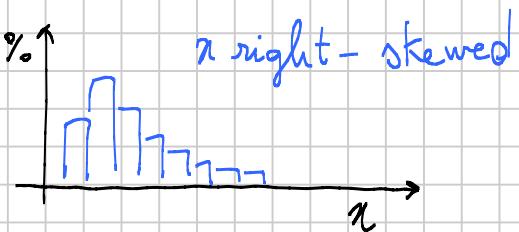
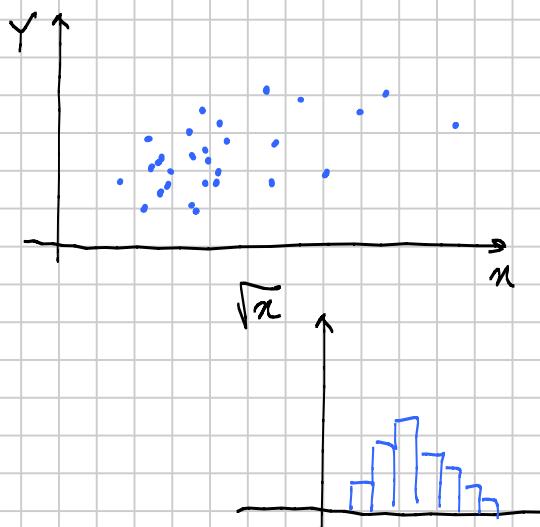
$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + e \quad \text{effetti additivi: } Y \text{ è di tipo "differente"}$$

* Trasformazioni lineari non cambiano la regressione

$$Y = \sum_{j=0}^p \beta_j x_j + N(0, \sigma^2) \quad \tilde{Y} = mY + q \quad \tilde{Y} = \sum_{j=1}^p (\tilde{m}\beta_j) x_j + (\tilde{m}\beta_0 + q) + N(0, \tilde{\sigma}^2)$$

\hookrightarrow i test, α^* , R_D^2 , R_A^2 non cambiano

* Trasformazioni nonlineari possono essere usate per migliorare la distribuzione di alcune variabili



Magari per la regressione aiuta

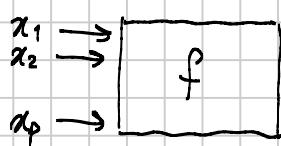
* In particolare è raccomandato di fare il logaritmo di una Y di tipo rapporto (lognormale)

$$Y: \text{stipendio} \quad x_1: \begin{cases} 1 & \text{frequentato Yale} \\ 0 & \text{no} \end{cases} \quad \leftarrow +40\% \quad \boxed{+10000\text{€}}$$

$$\log Y = \beta_0 + \beta_1 x_1 + \dots \quad Y = e^{\beta_0} \cdot \boxed{e^{\beta_1 x_1}} \cdot \dots \quad e^{\beta_1} = 1,4 \quad +40\%$$

■ REGRESSIONE LOGISTICA

ora 26



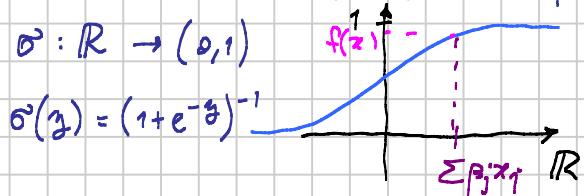
\rightarrow a volte categ. qualiasi:

$$(x_{ij}, z_i) \quad i=1, \dots, N, j=1, \dots, p$$

dicotomica : 0, 1 $Z \sim \text{bin}(n=1, p=f(x_1, x_2, \dots, x_p))$

$$f(x_1, \dots, x_p) = \delta(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)$$

$$\delta: \mathbb{R} \rightarrow (0, 1)$$



Si risolve massimizzando la likelihood

$$\begin{aligned} l(\beta_0, \beta_1, \dots, \beta_p) &= \log \left(\prod_{i=1}^N \varphi_{z_i}(x_i) \right) = \sum_{i=1}^N \log \left(\underbrace{\delta(\sum_j \beta_j x_{ij})}_{p_i}^{z_i} \cdot \underbrace{(1-\delta(\sum_j \beta_j x_{ij}))}_{1-p_i}^{1-z_i} \right) \\ &= \sum_{i=1}^N [z_i \log(p_i(x_i)) + (1-z_i) \log(1-p_i(x_i))] = -(\text{cross-entropy}) \end{aligned}$$

- * Non si sa la distribuzione dei β_j , quindi non si possono fare i test e la selezione delle variabili
- * Posso poi usare il modello addestrato per prendere la prob. di $z_i = \delta(\beta_0 + \dots + \beta_p x_p)$
- Versione multinomiale, con più di due categorie $k=1, \dots, m$
(vedi se 14, 15 e 16)

i	$(x(i))$	$Z(i)$	$\delta(B_1(i) \ B_2(i) \ B_3(i) \ B_4(i) \ \dots \ B_m(i))$
1	~	2	0 1 0 0 ... 0
2	~	1	1 0 0 0 ... 0
3	~	3	0 0 1 0 ... 0
:	~	1	1 0 0 0 ... 0
:	~	4	0 0 0 1 ... 0

$$x_{ij} = x_j(i) \quad i=1, 2, \dots, N \quad j=1, \dots, p \quad \text{dati di input}$$

$$z_i = z(i) \quad i=1, \dots, N \quad z_i \in \{1, 2, \dots, m\} \quad \text{output categorico}$$

$$B_{ik} = B_k(i) \quad i=1, \dots, N \quad k=1, \dots, m \quad \text{codifica one-hot}$$

$$B_{ik} := \mathbb{I}_{z_i=k}$$

parametri

modello : $B(i) \sim \text{multin}(1; \pi(x(i), w))$

$$\text{log-likel } \ell(w) = - \sum_{i=1}^N H(b(i); \pi(x(i), w)) \quad \text{vedi ora 16}$$

modello per π $\pi(x, w) := \text{sm}\left((w_0 + w_1 x_1 + w_2 x_2 + \dots + w_p x_p, \dots, w_0 + w_1 x_1 + w_2 x_2 + \dots + w_p x_p)\right)$

$$w \in M_{p+1, m}$$

parametri

$$y_k(i) = w_{0k} + w_{1,k} x_1(i) + \dots + w_{p,k} x_p(i)$$

logits

$$\pi_k(x(i), w) = [\text{sm}(y_k(i))]_k$$

probits

$$\frac{1}{N} \sum_{i=1}^N H(b(i); \pi(x(i), w))$$

loss

analogo alla regressione

con somma su k pari a 1

viene minimizzata

* Attenzione che non si fa mai la cross-entropy di sun

$$\frac{1}{N} \sum_{i=1}^N H_{\text{logit}}(b(i); y(i)) \quad \text{da usare in pratica}$$

* Quando uso il modello addestrato, trovo le probs delle m categorie (con somma 1) :

$$(\pi_1(i), \pi_2(i), \dots, \pi_m(i)) = \text{sm}(w_0 + x(i)w)$$

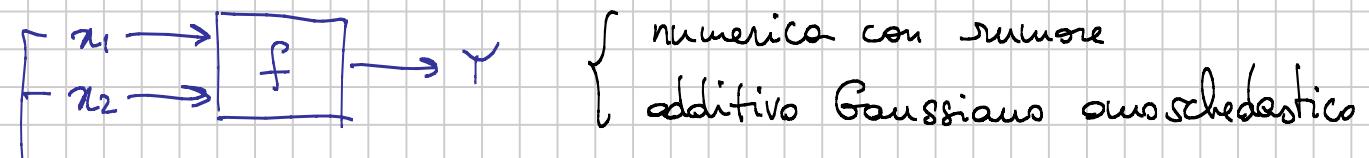
$w = (\underbrace{w_0, \dots, w_m}_\text{vettore}) \in \mathbb{R}^m$
 $x = (\underbrace{x_1, \dots, x_p}_\text{matrice}) \in M_{p, m}$

* Esiste anche la versione one-vs-many, in cui si fanno m regressioni logistiche dicotomiche, una per categoria

→ le probs trovate in predizione non sommano a 1

■ ANALISI DELLA VARIANZA

[capitolo 10 del Ross]

variante della regressione con variabili di ingresso categoricaliuna o due variabili categoriche (non troppe categorie)

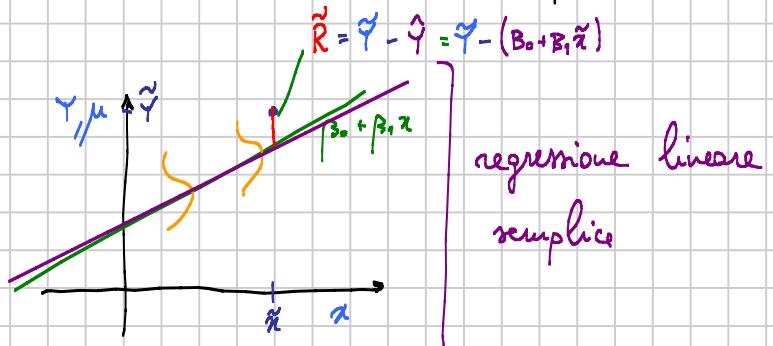
→ ANOVA a 1 VIA : una sola var. di ingresso

→ ANOVA a 2 VIE : due var. di ingresso (esattamente 1 osservazione per comb. di x_1, x_2)→ ANOVA a 2 VIE con repliche : $l \geq 2$ osservazioni per comb.

■ ANOVA A UNA VIA

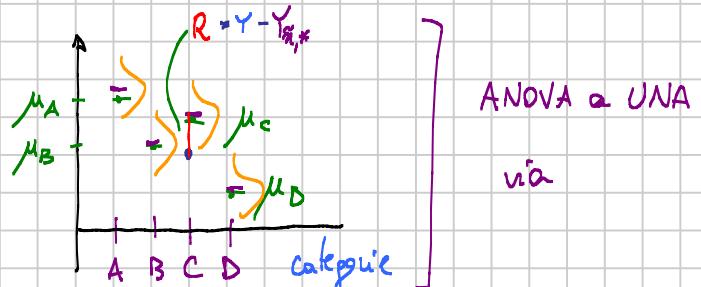
$$Y = \beta_0 + \beta_1 x + e$$

$$E(Y) = \mu = \mu(x) = \beta_0 + \beta_1 \cdot x$$

 β_0, β_1, σ params

$$Y = \mu_x + e$$

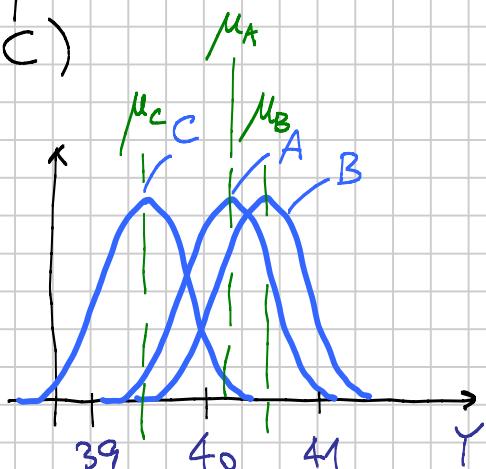
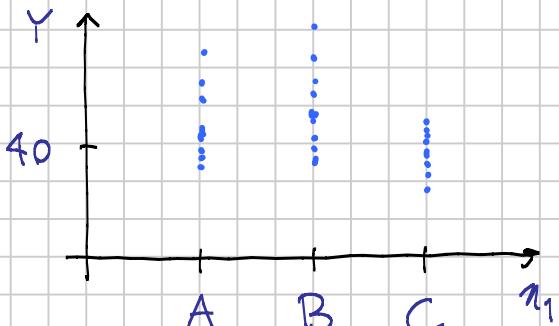
$$n \in \{A, B, C, D\}$$

 $\mu_A, \mu_B, \mu_C, \mu_D, \sigma$ params

→ Ogni categoria deve poter avere la propria media

→ esempio : x_1 ospedale (A, B, C)

Y : # settimane parto



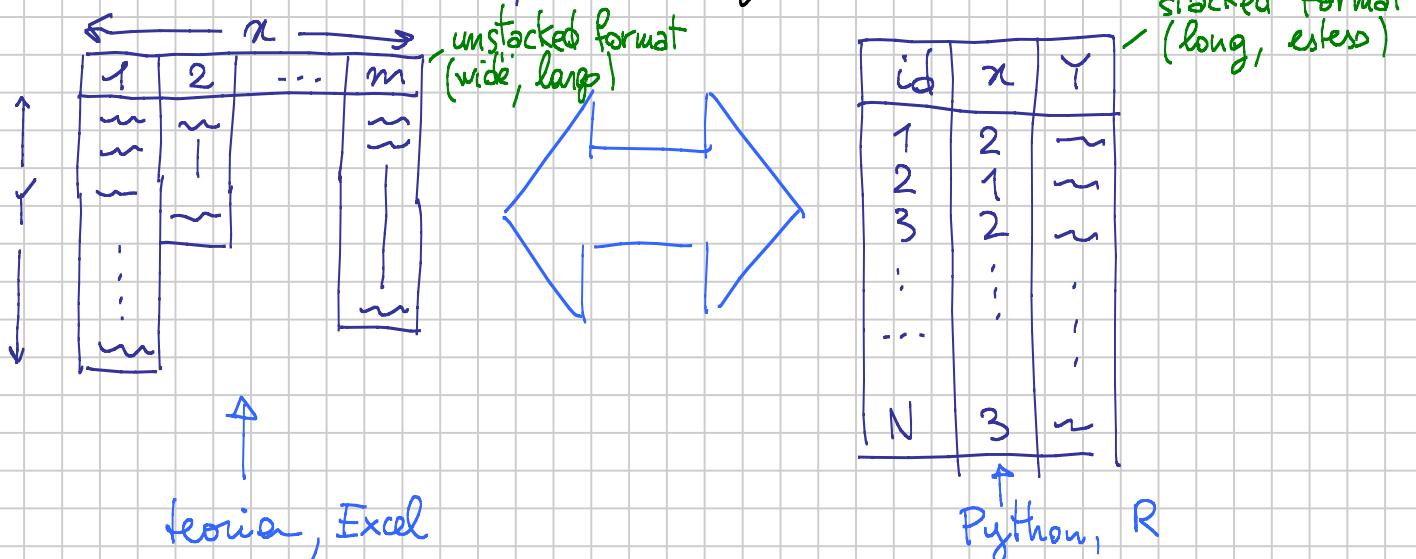
→ formalizzazione

$$x_i = x \in \{1, 2, \dots, m\} \quad \begin{matrix} \text{\# categorie} \\ (m \geq 2) \end{matrix} \quad \begin{matrix} \text{anche regressione} \\ m \geq 3 \end{matrix}$$

$$Y \sim N(\mu(x), \sigma^2) \quad \mu(x) = \mu_x \text{ medie eventualmente distinte}$$

i dati sono m campioni Gaussiani anche di numerosità diverse

→ due modi principali per raccogliere i dati



teoria, Excel

Y_{ij} : i la categoria ($=x$) j numero progressivo

$$Y_{ij} \quad i=1, 2, \dots, m \quad j=1, 2, \dots, n_i$$

$$\boxed{Y_{ij} \sim N(\mu_i; \sigma^2)} \quad \text{indipendenti} \quad \boxed{\# \text{dati categoria } i}$$

→ Parametri incogniti : $\mu_1, \mu_2, \dots, \mu_m, \sigma$

→ stimatori : $\mu_i \approx$ media campionaria categoria $i := \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} =: \bar{Y}_{i*}$

*: faccio la media su quell'indice

$$\text{distribuzione} : Y_{ij} \sim N(\mu_i; \sigma^2) \Rightarrow \frac{1}{n_i} \sum_j Y_{ij} \sim N\left(\frac{1}{n_i} \cdot n_i \mu_i; \frac{1}{n_i^2} \cdot n_i \sigma^2\right)$$

$$\Rightarrow \boxed{\bar{Y}_{i*} \sim N(\mu_i, \frac{\sigma^2}{n_i})} \quad i=1, 2, \dots, m$$

$\sigma^2 \approx$ varianza campionaria di ciascuna categoria: ho m stimatori!

$$S_i^2 := \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\bar{Y}_{ij} - \bar{Y}_{i*})^2 \approx \sigma^2 \quad (E(S_i^2) = \sigma^2)$$

per sintetizzare questi stimatori in uno singolo ne faccio la media:

$$S_A^2 := \frac{1}{m} \sum_{i=1}^m S_i^2 \approx \sigma^2 \quad (E(S_A^2) = \frac{1}{m} \sum_i E(S_i^2) = \frac{1}{m} \sum_i \sigma^2 = \frac{1}{m} \cdot m \sigma^2 = \sigma^2)$$

quando i campioni sono sbilanciati, S_A^2 non è ottimale: è rugoso e ha varianza elevata \rightarrow meglio pesare di più i campioni più grossi

$$S_T^2 := \sum_{i=1}^m \pi_i S_i^2 \approx \sigma^2 \quad \text{qui i } \pi_i \text{ sono dei pesi: } \pi_i \in [0, 1] \quad \sum_i \pi_i = 1$$

voglio π_i maggiore quando n_i è maggiore $(E(S_T^2) = \sigma^2)$

$$\text{Var}(S_T^2) = \sum_{i=1}^m \pi_i^2 \text{Var}(S_i^2) = \dots \Rightarrow \pi_i \text{ ottiene varianza minima con:}$$

$$\pi_i \propto \text{gdl}(S_i^2) = n_i - 1 \quad \text{HW: check (difficile)} \quad \left(\frac{S_i^2}{\sigma^2} (n_i - 1) \sim \chi^2(n_i - 1) \right)$$

$$\pi_i = \frac{n_i - 1}{\sum_{k=1}^m (n_k - 1)} = \frac{n_i - 1}{N - m} \quad \text{ho posto } N = n_1 + n_2 + \dots + n_m \quad (\text{dati totali})$$

- def Se ho m campioni casoschedastici indipendenti, ci scrivo con la mia media, lo stimatore within o pooled della varianza è

$$S_W^2 = S_p^2 = \sum_{i=1}^m \frac{n_i - 1}{N - m} S_i^2 = \frac{1}{N - m} \sum_{i=1}^m \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i*})^2 =: \frac{SS_W}{N - m} =: \frac{SS_R}{N - m}$$

La somma SS_W si chiama devianza within. \rightarrow residui dell'ANOVA

$$\rightarrow \frac{S_W^2}{\sigma^2} (N - m) \sim \chi^2(N - m) \quad \text{HW: check (facile)}$$

$$SS_W = \sum_{i,j} (Y_{ij} - \bar{Y}_{i*})^2 \quad \text{corrisponde a } SS_R = \sum_i (Y_i - (\sum_j B_j x_{ij}))^2$$

\uparrow somma dei quadrati dei residui

$$R_{ij} := Y_{ij} - \bar{Y}_{i*} \quad \text{residui dell'ANOVA a UNA via}$$

■ ANOVA A 1 VIA

✓ modello $Y_{ij} \sim N(\mu_i; \sigma^2)$ $j=1, 2, \dots, n_i$ $i=1, \dots, m$ $N = \sum_{i=1}^m n_i$

✓ parametri $\mu_1, \mu_2, \dots, \mu_m, \sigma$

✓ stime dei parametri $\hat{\mu}_i \approx Y_{i,*} \sim N(\mu_i; \frac{\sigma^2}{n_i})$ $\hat{\sigma} \approx S_w := \sqrt{\frac{1}{N-m} SS_w}$

$$SS_w := \sum_{i=1}^m (n_i - 1) S_i^2 = \sum_{i=1}^m \sum_{j=1}^{n_i} (Y_{ij} - Y_{i,*})^2$$

$$\frac{SS_w}{\sigma^2} \sim \chi^2(N-m)$$

$$\left| \frac{Y_{i,*} - \mu_i}{S_w} \sqrt{n_i} \sim t(N-m) \right|$$

f. anc. per μ_i f. anc. per σ

→ inferenza sui parametri ok: test $H_0: \sigma^2 \leq \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$

int. di conf per μ_i o di predizione per Y_{i,n_i+1}

HW: think!

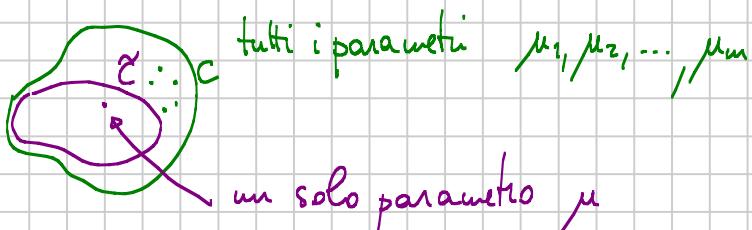
② Test fondamentale della ANOVA a una via

$$H_0: \underbrace{\mu_1 = \mu_2 = \dots = \mu_m}_{Y \text{ non dipende da } x}$$

$$H_1: \text{non tutte uguali}$$

\uparrow
 Y dipende da x

(vedi no 23) come allora, voglio confrontare due modelli



posso "fittare" i miei dati con un modello **a un solo parametro** o con un modello **a m parametri**

i. $Y_{ij} \sim N(\mu, \sigma^2)$ $\mu \approx \bar{Y} = \frac{1}{N} \sum_{i=1}^m \sum_{j=1}^{n_i} Y_{ij}$

"residui" sono $T_{ij} - \bar{Y}$ $SS_Y = \sum_i \sum_j (Y_{ij} - \bar{Y})^2$

ii. $Y_{ij} \sim N(\mu_i, \sigma^2)$ ANOVA a 1 via $\mu_i \approx Y_{i,*}$

"residui" sono $Y_{ij} - Y_{i,*}$ $SS_w = \sum_i \sum_j (Y_{ij} - Y_{i,*})^2$

$$SS_B := SS_Y - SS_W$$

$m-1$ $N-1$ $N-m$
g.d.l. g.d.l. g.d.l.

$$V := \frac{SS_B / (m-1)}{SS_W / (N-m)}$$

statistica del test

- Test anova a 1 via : calcolo gli stimatori *within* già visti e ne ricavo gli stimatori *between*

$$SS_B = SS_Y - SS_W$$

$$SS_B = \sum_{i=1}^m n_i (Y_{i,*} - \bar{Y})^2$$

devianza "tra un campione e l'altro"

🚩 $S_B^2 := \frac{SS_B}{m-1}$ varianza *between* :

sotto H_0 c'è poi pure uno stimatore corretto di σ^2

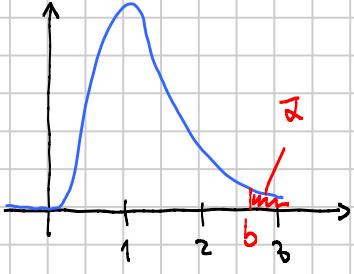
sotto H_1 è grande

va divisa per $S_W^2 \approx \sigma^2$ per poter valutare se è grande o no

→ Statistica del test :

🚩 $F_{\text{ANOVA}} = \frac{S_B^2}{S_W^2} = \frac{SS_B / (m-1)}{SS_W / (N-m)} \stackrel{H_0}{\sim} F(m-1; N-m)$

Sotto H_1 F_{ANOVA} assume valori grandi \rightarrow test unilaterale



$$b = F_{F(\dots)}^{-1}(1-\alpha)$$

$$RA_{F_{\text{ANOVA}}} = [0; b]$$

$$\alpha^* = 1 - F_{F(\dots)}(F_{\text{ANOVA}})$$

- Verifica dell'identità delle devianze

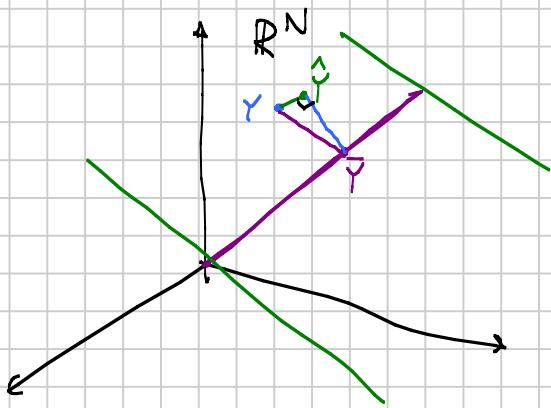
$$Y = (Y_{ij})_{ij} = (Y_{11}, Y_{12}, \dots, Y_{1n}, Y_{21}, \dots) \in \mathbb{R}^N$$

$$Y \approx \bar{Y} := (\bar{Y}, \bar{Y}, \dots, \bar{Y}, \bar{Y}, \dots) \in \mathbb{R}^N$$

$$Y_{ij} \sim \mathcal{N}(\mu, \sigma^2)$$

$$Y \approx \hat{Y} := (Y_{1*}, Y_{2*}, \dots, Y_{1*}, Y_{2*}, \dots) \in \mathbb{R}^N$$

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$$



$$\begin{array}{ll} | & Y - \bar{Y} \quad Y_{ij} - Y_{ij*} \\ | & \bar{Y} - \bar{Y} \quad Y_{ij*} - \bar{Y} \\ | & Y - \bar{Y} \quad Y_{ij} - \bar{Y} \end{array}$$

$$\sum_i \sum_j (\gamma_{ij} - \bar{\gamma}_{ix})(\gamma_{ijx} - \bar{\gamma}) = \sum_i (\bar{\gamma}_{ix} - \bar{\gamma}) \sum_{j=1}^{n_x} (\gamma_{ij} - \bar{\gamma}_{ix}) = 0$$

$$\text{quindi: } \sum_{i=1}^n \sum_{j=1}^{k_i} (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^n \sum_{j=1}^{k_i} (Y_{ij} - \bar{Y}_{ij})^2 + \sum_{i=1}^n \sum_{j=1}^{k_i} (\bar{Y}_{ij} - \bar{Y})^2$$

SS_Y SS_W SS_B

$$SS_B = \sum_i \sum_j (Y_{ij} - \bar{Y})^2 = \sum_i (Y_{i*} - \bar{Y})^2 \cdot n_i$$

• Attenzione alle cose pratiche :

→ residui : no dipendenza da x , Gaussiani, omosched.
no outliers
(grafici)

→ se serve trasformazioni nonlineari per γ

→ se le variabili di ingresso non consentono l'ANOVA:
regressione con categorie gestite.

→ se è accettabile H_0 , può essere il caso di ignorare questa χ e studiare i dati come campione Gaussiano iid $Y_i \sim N(\mu, \sigma^2)$

⑥ Approfondimenti sul Ross

confronti multipli fra medie

$$\text{sotto } H_1 \quad E(S_B^2) > \sigma^2$$

$$E\left[\frac{SS_b}{m-1}\right] = \sigma^2 + \frac{n}{m-1} \sum_{i=1}^m (\mu_i - \mu_*)^2$$

Proposizione 10.3.2. Per ogni scelta degli indici i, j diversi tra loro, e per ogni $\alpha \in (0, 1)$, con probabilità $1 - \alpha$,

$$X_{i*} - X_{j*} - W < \mu_i - \mu_j < X_{i*} - X_{j*} + W \quad (10.3.14)$$

dove si è posto

$$W := \frac{1}{\sqrt{nm}} C(m, nm - m, \alpha) \sqrt{SS_W / (nm - m)} \quad (10.3.15)$$

I valori dei coefficienti $C(m, d, \alpha)$ per $\alpha = 0.01$ e $\alpha = 0.05$ sono riportati nella Tabella A.5 in Appendice. In R sono disponibili tramite la funzione `qtukey()`, in particolare, se `alpha` contiene il valore di α , allora

$$C(m, d, \alpha) = \text{qtukey}(1 - \alpha, m, d)$$

INTRODUZIONE AL MACHINE LEARNING

Note Title

ora 28

06/05/2025

■ ANOVA A DUE VIE (anche con REPLICHE)

$$x_1 \in \{1, 2, \dots, m\}$$

$$x_2 \in \{1, 2, \dots, n\}$$

ℓ : # di repliche, $\ell \geq 1$

		unstacked			
		1	2	...	n
1		~~~	~~~	...	~~~
		~~~	~~~	...	~~~
2		~~~	~~~	...	~~~
		~~~	~~~	...	~~~
:	
m		~~~	~~~	...	~~~
		~~~	~~~	...	~~~

$\} \ell$  repliche

$\} \ell$  repliche

$\} \ell$  repliche

stacked			
id	$x_1$	$x_2$	$y$
1			
2			
:			
N			

$$N = m \cdot n \cdot \ell$$

formato strumenti  
di analisi Excel

se è così, occorre contare le  
repliche  $\ell_{ij}$  per ogni combinaz  
e vanno gestite le cose in modo da  
ottenere il costante  
( buttando via dati e/o categorie,  
ma il meno possibile

### • Notazione e modelli

$$Y_{ijk} \quad \underbrace{i=1, 2, \dots, m}_{x_1} \quad \underbrace{j=1, 2, \dots, n}_{x_2} \quad \underbrace{k=1, 2, \dots, \ell}_{\text{repliche}}$$

$$Y_{ijk} \sim \mathcal{N}(\mu_{ijk}; \sigma^2)$$

modello generalissimo: ogni casella ha la  
sua media specifica

Questo modello ha  $m \cdot n + 1$  parametri ( $\mu_{11}, \mu_{12}, \dots, \mu_{mn}, \sigma$ ): stimabili solo se  $\ell \geq 2$

Se possibile, si cerca di ridursi ad un modello additivo molto più semplice

$$M_{ij} = \mu + \alpha_i + \beta_j$$

modello della ANOVA a 2 vie senza r.

si impone  $\sum_i \alpha_i = 0 = \sum_j \beta_j$       m+n ven

→ questo modello ha  $m+n+2$  parametri

Per capire se è possibile, per prima cosa si riscrive  $M_{ij}$  come:

$$M_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

modello della ANOVA a 2 vie con repliche

$\sum_i \alpha_i = 0 = \sum_j \beta_j$      $\sum_i \gamma_{ij} = 0 = \sum_j \gamma_{ij}$       m.n+1 pars

termine di interazione

→ Questa riscrittura è sempre possibile, ed è unica perché si impone che i vari effetti abbiano media nulla:

$$\alpha_* = \beta_* = \gamma_{*,j} = \gamma_{i,*} = 0$$

HW: check

→ Esiste un test per verificare  $H_0: \gamma_{ij} = 0$      $H_1: \text{non tutti nulli}$

i. se viene  $H_0$  è corretto usare il modello additivo = ANOVA a 2 vie SENZA r.

↪ per fare "sparire" le repliche e scendere l a 1 si sostituiscono

i dati originali  $Y_{ijk}$  con le medie  $Y_{ij,*}$

	1	2	...	n
1	~	~	...	~
2	~	~	...	~
⋮	...	...	...	...
m	~	~	...	~



	1	2	...	n
1	~	~	...	~
2	~	~	...	~
⋮	...	...	...	...
m	~	~	...	~

ii. se viene  $H_1$  si deve tenere il modello generalissimo, che è più difficile da interpretare

## Test fondamentale ANOVA a 2 vie

→ sono due test, per le righe e per le colonne

→ si fanno sia nel caso con repliche, sia in quello senza

$H_0: Y$  non dipende da  $x_1$

"non c'è effetto riga"

$$\alpha_1 = \alpha_2 = \dots = \alpha_m = 0$$

$H_1: Y$  dipende da  $x_1$

"c'è effetto riga"

$\alpha_i$  non tutti nulli

→ è analogo per  $x_2$  e le colonne

Test in pratica, dal Ross :

Tabella 10.3 ANOVA a due fattori.

CASO SENZA REPLICHE

	Somma di quadrati	Gradi di libertà
Riga	$SS_B \sim \left\{ \begin{array}{l} SS_r := n \sum_i (X_{i*} - X_{**})^2 \\ SS_c := m \sum_j (X_{*j} - X_{**})^2 \end{array} \right.$	$m - 1$
Colonna	$SS_e := \sum_i \sum_j (X_{ij} - \bar{X}_{ij} \text{ previsto})^2$	$n - 1$
Errore	$SS_e := \sum_i \sum_j (X_{ij} - \bar{X}_{ij} \text{ previsto})^2$	$(n - 1)(m - 1)$

non è il numero  
di dati → Sia  $N = (n - 1)(m - 1)$ .

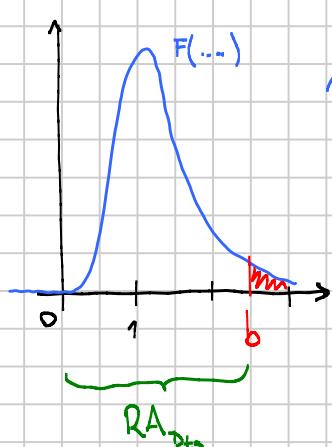
Ipotesi nulla	Statistica del test	Un test con significatività $\alpha$ deve rifiutare $H_0$ se $D_{ts} > F_{\alpha, m-1, N}$	$p$ -dei-dati se $D_{ts} = v$
Tutte le $\alpha_i = 0$	$D_{ts} := \frac{SS_r}{SS_e} (n - 1)$	rifiutare $H_0$ se $D_{ts} > F_{\alpha, m-1, N}$	$P(F_{m-1, N} \geq v)$
Tutte le $\beta_j = 0$	$D_{ts} := \frac{SS_c}{SS_e} (m - 1)$	rifiutare $H_0$ se $D_{ts} > F_{\alpha, n-1, N}$	$P(F_{n-1, N} \geq v)$

$$\frac{SS_r}{\sigma^2} \stackrel{H_0}{\sim} \chi^2(m-1) \rightarrow \boxed{\frac{SS_r}{m-1} = S_r^2 \approx \sigma^2}$$

$$D_{ts} := \frac{S_r^2}{S_e^2} = \frac{SS_r}{m-1} \cdot \frac{(n-1)(m-1)}{SS_e} = \frac{SS_r}{SS_e} (n-1)$$

$$\frac{SS_e}{\sigma^2} \sim \chi^2((n-1)(m-1)) \rightarrow \boxed{\frac{SS_e}{(n-1)(m-1)} = S_e^2 \approx \sigma^2}$$

corretto sempre



$$\text{notto } H_0 \left\{ \begin{array}{l} D_{ts} \sim F(m-1, (m-1)(n-1)) \\ D_{ts} \sim F(n-1, (m-1)(n-1)) \end{array} \right.$$

test sull'effetto riga  
test sull'effetto colonna

## • Stimatori (caso senza repliche)

$$Y_{ij} \sim N(\mu + \alpha_i + \beta_j; \sigma^2)$$

$$Y_{i*} = \frac{1}{n} \sum_{j=1}^n Y_{ij} \sim N\left(\underbrace{\frac{1}{n} \sum_{j=1}^n (\mu + \alpha_i + \beta_j)}_{= \mu + \alpha_i + \beta_*}; \frac{\sigma^2}{n}\right) \sim N(\mu + \alpha_i; \frac{\sigma^2}{n})$$

$$Y_{*j} \sim N(\mu + \beta_j; \frac{\sigma^2}{m}) \quad Y_{**} \sim N(\mu; \frac{\sigma^2}{mn})$$

$$\alpha_i \approx Y_{i*} - Y_{**} \sim N(\alpha_i; \text{mostro})$$

$$\beta_j \approx Y_{*j} - Y_{**} \sim N(\beta_j; \text{altro mostro})$$

HW: calcolare varianze

previsto  $Y_{ij}$  aka. stima di  $E(Y_{ij}) = \mu + \alpha_i + \beta_j \approx Y_{i*} + Y_{*j} - Y_{**}$

residui:  $R_{ij} = Y_{ij} - Y_{i*} - Y_{*j} + Y_{**}$

$$SS_e := \sum_{ij} R_{ij}^2 \quad \text{devianza d'errore}$$

$$S_e := \sqrt{\frac{SS_e}{(m-1)(n-1)}} \approx \sigma$$

Lo stimatore di  $\sigma^2$  che è sempre corretto, indipendentemente dalle ipotesi è

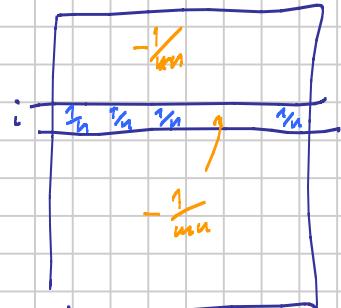
$$\hat{\sigma}^2 \approx S_e^2 = \frac{SS_e}{(m-1)(n-1)}$$

$$\frac{SS_e}{\sigma^2} \sim \chi^2_{((m-1)(n-1))}$$

→ Faccio il conto della varianza di  $Y_{i*} - Y_{**}$

$$\begin{aligned} Y_{i*} - Y_{**} &= \frac{1}{n} \sum_{j=1}^n Y_{ij} - \frac{1}{mn} \sum_{k=1}^m \sum_{j=1}^n Y_{kj} \\ &= \sum_{k \neq i} -\frac{1}{mn} \sum_{j=1}^n Y_{kj} + \left(\frac{1}{n} - \frac{1}{mn}\right) \sum_{j=1}^n Y_{ij} \\ &= -\frac{1}{mn} \underbrace{\sum_{k \neq i} \sum_{j=1}^n Y_{kj}}_{n \cdot (m-1)} + \frac{m-1}{mn} \underbrace{\sum_{j=1}^n Y_{ij}}_n \end{aligned}$$

ma sono tutte  
indipendenti



$$\text{Var}(Y_{i*} - Y_{**}) = \left(\frac{1}{mn}\right)^2 \cdot n \cdot (m-1) \cdot \sigma^2 + \frac{(m-1)^2}{m^2 n^2} \cdot n \cdot \sigma^2 = \frac{m-1}{m^2 n} \sigma^2 (1 + m-1) = \frac{m-1}{mn} \sigma^2$$

## • Stimatori (caso con repliche)

$$Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$$

$$\mu_{ij} \approx Y_{ij*} = \frac{1}{e} \sum_{k=1}^e Y_{ijk} \sim N(\mu_{ij}; \frac{\sigma^2}{e}) \quad \text{previsto}$$

$$R_{ijk} = Y_{ijk} - Y_{ij*} \quad \text{residui}$$

$$SS_e = \sum_{i,j,k} R_{ijk}^2$$

$$Y_{ijk} = \frac{1}{n \cdot l} \sum_{j,k} Y_{ijk} \sim N(\mu + \alpha_i + \beta_j; \frac{\sigma^2}{n \cdot l}) \quad x_i \approx Y_{ij*} - Y_{*jk*} \sim N(x_i; \dots)$$

$$\mu + \alpha_i + \beta_j \approx Y_{ij*} + Y_{*jk*} - Y_{*ik*} \sim N(\mu + \alpha_i + \beta_j; \dots)$$

$$Y_{ij} = \mu_{ij} - (\mu + \alpha_i + \beta_j) \approx Y_{ij*} - Y_{*jk*} - Y_{*ik*} + Y_{*kk*}$$

### CASO CON REPLICHE

**Tabella 10.4** ANOVA a due fattori, con interazioni e  $l$  osservazioni per cella. Si è posto  $N := mn(l-1)$  e  $M := (n-1)(m-1)$ .

Fonte di variabilità	Somma di quadrati	Gradi di libertà
Riga	$SS_r := nl \sum_{i=1}^m (X_{i**} - X_{***})^2$	$m-1$
Colonna	$SS_c := ml \sum_{j=1}^n (X_{*j*} - X_{***})^2$	$n-1$
Interazioni	$SS_{int} := l \sum_{i=1}^m \sum_{j=1}^n (X_{ij*} - X_{i**} - X_{*j*} + X_{***})^2$	$M$
Errore	$SS_e := \sum_{k=1}^l \sum_{i=1}^m \sum_{j=1}^n (X_{ijk} - X_{ij*})^2$	$N$
Ipotesi nulla	Statistica del test	Un test con significatività $\alpha$ deve se $F = v$
$H_0^r$ : Le $\alpha_i$ sono tutte nulle	$F_r := \frac{SS_r / (m-1)}{SS_e / N}$	rifiutare $H_0^r$ se $F_r > F_{\alpha, m-1, N}$ $P(F_{m-1, N} \geq v)$
$H_0^c$ : Le $\beta_j$ sono tutte nulle	$F_c := \frac{SS_c / (n-1)}{SS_e / N}$	rifiutare $H_0^c$ se $F_c > F_{\alpha, n-1, N}$ $P(F_{n-1, N} \geq v)$
$H_0^{int}$ : Le $\gamma_{ij}$ sono tutte nulle	$F_{int} := \frac{SS_{int} / M}{SS_e / N}$	rifiutare $H_0^{int}$ se $F_{int} > F_{\alpha, M, N}$ $P(F_{M, N} \geq v)$

questo determina se usare il modello additivo o generale sotto  $H_0$

$$\begin{cases} F_r \sim F(m-1, mn(l-1)) & \text{test sull'effetto riga} \\ F_c \sim F(n-1, mn(l-1)) & \text{test sull'effetto colonna} \\ F_{int} \sim F((m-1)(n-1), mn(l-1)) & \text{test sulle interazioni} \end{cases}$$

Lo stimatore di  $\sigma^2$  che è sempre corretto, indipendentemente dalle ipotesi è

$$\hat{\sigma}^2 \approx S_e^2 = \frac{SS_e}{mn(l-1)}$$

$$\frac{SS_e}{\hat{\sigma}^2} \sim \chi^2(mn(l-1))$$

→ se l'ento del test sull'effetto di una variabile (riga o colonna) è  $\epsilon H_0$ , si può considerare di eliminare quella variabile e ridursi all'ANOVA a 1 via

## TEST DEL CHI-QUADRO

- sono tre-quattro famiglie di test di adattamento ad una distribuzione
  - i. test del chi-quadro elementare : per distribuzioni discrete con pochi valori  
↳ ad esempio : distribuzione teorica del dado o d'oro
  - ii. generalizzazione a distribuzioni qualitativi : con tanti valori o continue  
↳ ad esempio : distribuzione delle estazioni del lotto  
: distribuzione dei numeri pseudocasuali di rand()
  - iii. generalizzazione a famiglia di distribuzioni : a meno di parametri  
↳ ad esempio : distribuzione Gaussiana ( $\mu$  e  $\sigma$  qualitativi)
  - iv. tabelle di contingenza : verifica se la distribuzione congiunta di due variabili è quella delle variabili indipendenti

## TEST DEL CHI-QUADRO ELEMENTARE

campione  $X_1, \dots, X_n$  con distribuzione incognita

- i. la distribuzione incognita  $\varphi$  è una legge discreta che assume un numero modesto  $k$  di valori possibili
- ii. c'è una distribuzione "candidata privilegiata"  $\varphi_0$  tra tutte quelle possibili (ad esempio la distrib. del dado o d'oro) e vogliamo verificare se effettivamente è quella vera :

$$H_0: \varphi = \varphi_0$$

$$H_1: \varphi \neq \varphi_0$$

→ esempio dado :  $\varphi$  potrebbe essere  $(0.1, 0.1, 0.2, 0.2, 0.3, 0.1)$   
o qualunque altra

$$p_0 = \left( \frac{1}{6}, \frac{1}{6}, \dots, \frac{1}{6} \right)$$

$$\Delta_6 = \{(p_1, p_2, \dots, p_6) \geq 0 : p_1 + \dots + p_6 = 1\}$$

→ operativamente : conta quante volte nel campione  $x_1, \dots, x_n$  è uscito ciascuno dei  $k$  valori :  $\{1, 2, \dots, k\}$

$O_1$ : # di volte che è uscito 1

$O_2$ : # di volte che è uscito 2

...

$$j=1, 2, \dots, k \quad O_j := \#\{i=1, 2, \dots, n : x_i = j\} := \sum_{i=1}^n \mathbb{1}_{x_i=j}$$

osservati

→ distribuzione degli osservati :  $(O_1, O_2, \dots, O_k) \sim \text{multin}(n, \varphi)$

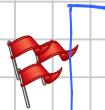
in particolare :  $O_j \sim \text{bin}(n, \varphi(j)) \Rightarrow E(O_j) = n\varphi(j)$

$$j=1, 2, \dots, k \quad A_j := n\varphi(j) = E_{H_0}(O_j) = E_{H_0}(O_j)$$

attesi

→ statistica del test : misura in scala opportuna lo scostamento tra attesi e osservati

$$O_1 + O_2 + \dots + O_k = n = A_1 + A_2 + \dots + A_k$$



$$W := \sum_{j=1}^k \frac{(O_j - A_j)^2}{A_j}$$

statistica del test (Pearson)

↪ grande sotto  $H_1$ , "piccola" sotto  $H_0$

→ distribuzione di  $W$

$$W \stackrel{H_0}{\sim} \chi^2(k-1)$$



è un risultato asintotico vero per  $n \rightarrow \infty$



* "rule of thumb" :  $A_j \geq 1$  tutti  $A_j \geq 5$  almeno l'80%

→ negli ultimi anni si tende a preferire la statistica  $G$

$$G = 2 \sum_{j=1}^k O_j \log \left( \frac{O_j}{A_j} \right) \geq 0$$

statistica del test  $G$

$$G \stackrel{H_0}{\sim} \chi^2(k-1)$$

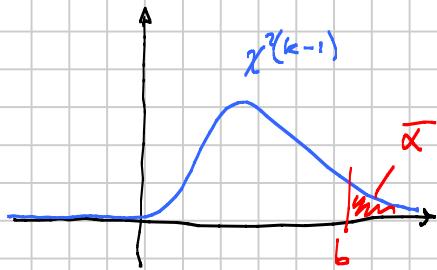
l'approssimazione dovrebbe essere migliore

→ test unilaterale :

$$RA_W = RA_G = [0; b]$$

$$b = F_{\chi^2(k-1)}^{-1}(1-\alpha)$$

$$\boxed{\alpha^* = 1 - F_{\chi^2(k-1)}(W \circ G)}$$



→ E' un test poco potente : serve  $A_i \gg 1$  per rilevare differenze modeste tra  $\varphi$  e  $\varphi_0$ .

### • Una derivazione della statistica W di Pearson

$$O = (O_1, O_2, \dots, O_k) \sim \text{multin}(n, \varphi)$$

$$P(O_1=j_1, O_2=j_2, \dots, O_k=j_k) = \frac{n!}{j_1! j_2! \dots j_k!} \varphi_1^{j_1} \cdot \varphi_2^{j_2} \cdots \varphi_k^{j_k} \quad j_1 + j_2 + \dots + j_k = n$$

(complicata, componenti legate)

Alternativa :

$$\tilde{O}_1, \tilde{O}_2, \dots, \tilde{O}_k \text{ iudip} \quad \tilde{O}_i \sim \text{Pois}(n\varphi_i) \quad S = \tilde{O}_1 + \tilde{O}_2 + \dots + \tilde{O}_k$$

$$P(\tilde{O}_1=j_1, \tilde{O}_2=j_2, \dots, \tilde{O}_k=j_k) = \prod_{i=1}^k \frac{(n\varphi_i)^{j_i} e^{-n\varphi_i}}{j_i!} = \frac{n^{j_1+j_2+\dots+j_k}}{j_1! j_2! \dots j_k!} \varphi_1^{j_1} \cdot \varphi_2^{j_2} \cdots \varphi_k^{j_k} \cdot e^{-n} \quad j_1, \dots, j_k \geq 0$$

(molto simile!)

* La legge di  $\tilde{O} = (\tilde{O}_1, \dots, \tilde{O}_k)$  condizionata a  $S=n$  è uguale a quella di  $O$

$$\text{HW: } P(\tilde{O}_1=j_1, \tilde{O}_2=j_2, \dots, \tilde{O}_k=j_k \mid S=n) := \frac{P(\tilde{O}_1=j_1, \tilde{O}_2=j_2, \dots, \tilde{O}_k=j_k, S=n)}{P(S=n)} = P(O_1=j_1, O_2=j_2, \dots, O_k=j_k)$$

$$E(\tilde{O}_i) = n\varphi_i = \text{Var}(\tilde{O}_i) =: \mu_i$$

$$\text{TLC: } n \gg 1 \quad \tilde{O}_i \sim N(\mu_i, \mu_i)$$

$\frac{\tilde{O}_i - \mu_i}{\sqrt{\mu_i}}$  ha media 0 e var 1

$$\xrightarrow{\text{~}} n \mathcal{N}(0, 1) \rightarrow \frac{(\tilde{O}_i - \mu_i)^2}{\mu_i} \xrightarrow{\text{~}} \chi^2(1)$$

$$\text{Sotto } H_0 : \varphi = \varphi_0 \quad \mu_i = n\varphi_i = n\varphi_0 = A_i \quad \rightarrow \quad \frac{(\tilde{O}_i - A_i)^2}{A_i} \stackrel{H_0}{\sim} \chi^2(1)$$

$$\rightarrow \sum_{i=1}^k \frac{(\tilde{O}_i - A_i)^2}{A_i} \stackrel{H_0}{\sim} \chi^2(k)$$

Condizionare a  $S=n$  in qualche modo cala di 1 i gdf

## ESTENSIONE A LEGGI QUALSIASI

- leggi discrete con tanti valori possibili (anche  $\infty$ )
- leggi continue

$\varphi_0$ : legge compatibile, completamente specificata

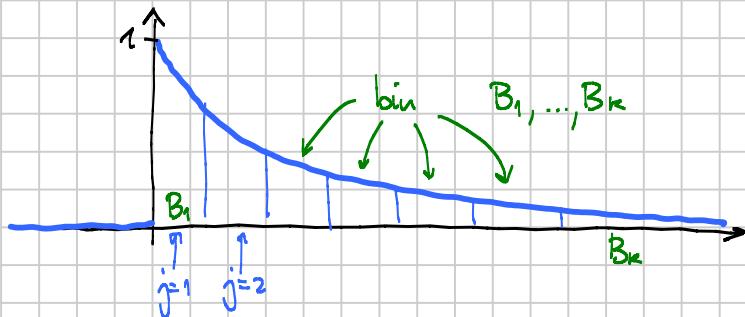
Fisso  $k$ , divido  $\mathbb{R}$  in  $k$  insiemi (bin) e poi conta quanti dati cadono in ciascun insieme

→ esempio:  $F_0 \sim \text{expo}(1)$

$H_0: F \sim F_0$

$$k=7 \quad j=1, \dots, k$$

$$A_j = n P_{H_0}(X_1 \in B_j) = n \int_{B_j} f_0(t) dt$$

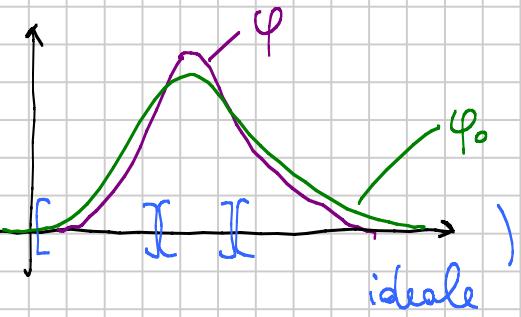
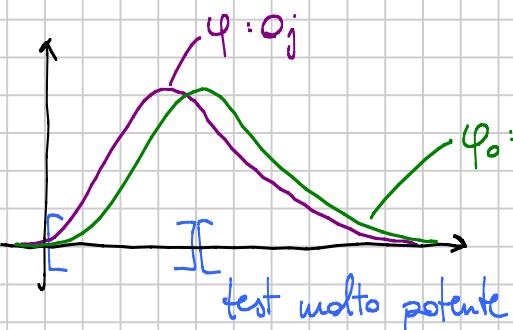


→ Attenzione: il test è più potente e l'approssimazione è migliore, a parità di  $k$  se gli  $A_j$  sono più grandi possibile → bin equiprobabili, non larghi uguali.

→ La capacità di dire  $H_1$  (potenza) è comunque legata al numero e alle posizioni dei bin

vera  $H_1: \varphi \neq \varphi_0$

dall'ora 30 del 2017



[ ] [ ] [ ] [ ] [ ] )  
test non molto potente

[ ] [ ] )  
permesso

## ■ ESTENSIONE A LEGGE SPECIFICATA A MENO DI PARAMETRI

esempio:  $H_0: X \sim N$        $H_1: X \not\sim N$

(si può usare come test di Gaussianità)

→ stimo  $\mu \approx \bar{x}$ ,  $\sigma \approx S_x$  dal campione

→ faccio il test  $H_0: X \sim N(\bar{x}, s^2)$

(come in precedenza)

↳ unica differenza, dovò togliere un g.d.l.  
per ogni parametro stimato

$$W \stackrel{H_0}{\sim} \chi^2(k-3) \quad (\text{nella' esempio } W)$$

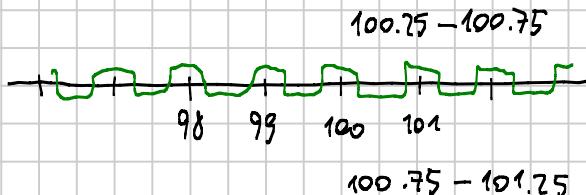
### ② Sui test di gaussianità e simili

→ A volte serve verificare la gaussianità a priori o a posteriori dell'uso di qualche tecnica statistica: in questi casi fare il test **non è una buona idea**. La risposta è un **trade-off** fra vicinanza con la **distribuzione gaussiana** e **numerosità del campione**

L'ipotesi vera che vorrei è  $H_0: d(x, N) < \text{qualcosa}$

perché una distribuzione vicina a  $N$  mi va bene

Il test non fa e non può fare questo: **meglio fare anche un controllo qualitativo**



esempio / provocatione

## ■ TABELLE DI CONTINGENZA

servono a testare se due variabili sono indipendenti ( $H_0$ ) o legate da una relazione ( $H_1$ )

esempio: patologia → due cure da testare → efficacia dicotomica

↳ se  $H_0$ : le cure non sono efficaci

guarig	ctrl	cure1	cure2	
si	7	13 ¹⁰	20 ¹¹	40
no	17	20 ²³	17 ²⁶	54
	24	33	37	94

Il test deve verificare se "qualche" trattamento ha efficacia diversa, nel senso di una diversa proporzionalità fra le colonne  
→ equivalentemente, fra le righe

$O_{ij} \quad i=1,2 \quad j=1,2,3$  sono gli osservati

guarig	ctrl	cure1	cure2	
si	10.2	14.0	15.8 ⁴	40
no	13.8	19.0	21.2 ⁵	54
	24	33	37	94

$A_{ij} \quad i=1,2 \quad j=1,2,3$  sono gli attesi

$$A_{ij} = \frac{(\text{totale riga } i) \cdot (\text{totale colonna } j)}{N}$$

$$W = \sum_{i,j} \frac{(O_{ij} - A_{ij})^2}{A_{ij}} \stackrel{H_0}{\sim} \chi^2((m-1)(n-1))$$



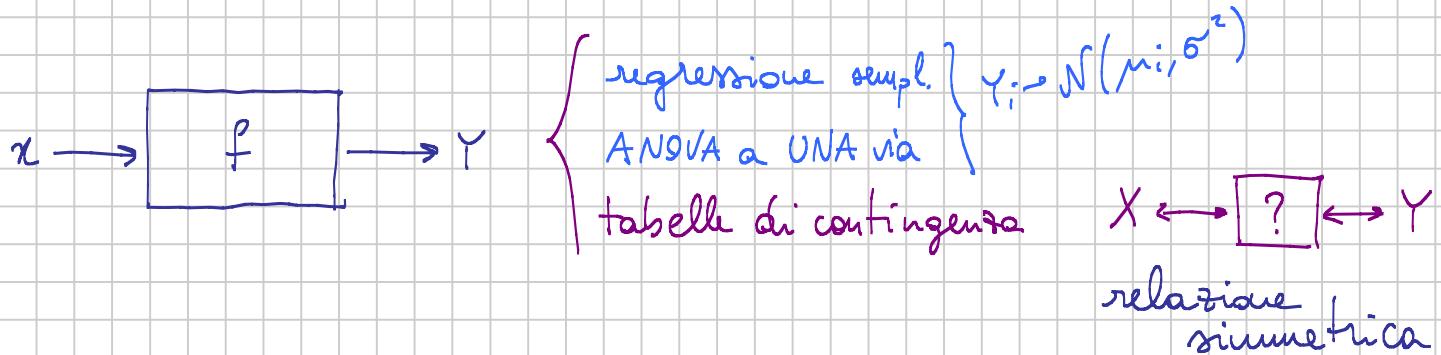
* solita "rule of thumb"  
80%  $A_{ij} \geq 5$  100%  $A_{ij} \geq 1$

$m, n$ : # di righe e di colonne

* Anche

$$G = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{A_{ij}} \stackrel{H_0}{\sim} \chi^2((m-1)(n-1))$$

* Vari modi per studiare se due variabili sono legate/correlate

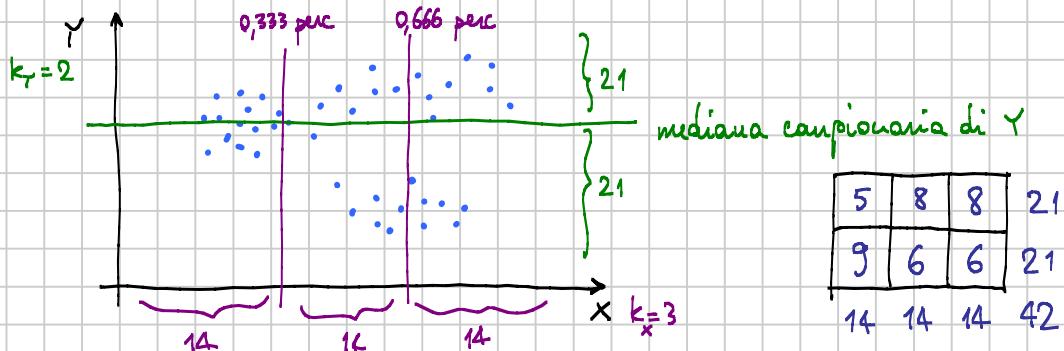


* tabelle di contingenza: di base quando le variabili sono discrete con pochi(ssimi) valori.

* sono un test non parametrico (no ipotesi sulla famiglia della distrib.)  
 ↳ meno potenti di ANOVA e regr. lin.

* se le due variabili (diciamo  $X$  e  $Y$ ) hanno tanti valori:

↳ dividere  $X$  e  $Y$  in bin cercando di avere marginali equidistribuiti



### Sketch della teoria

dati :  $(x_k, Y_k)$   $k = 1, 2, \dots, N$

$$x_k \in \{1, 2, \dots, m\} \quad Y_k \in \{1, 2, \dots, n\}$$

$H_0$ :  $X, Y$  indipendenti

$$\varphi_{x,y}(i,j) = \varphi_x(i) \cdot \varphi_y(j) \quad \text{relazione di indipendenza}$$

- osservati :  $\forall i, j \quad i \in 1, \dots, m \quad j \in 1, \dots, n$

$$O_{ij} := \#\{k \in 1, \dots, N : X_k = i, Y_k = j\} \sim \text{bin}(N, \varphi_{x,y}(i,j)) \quad E(O_{ij}) = N \varphi_{x,y}(i,j)$$

- attesi :  $A_{ij}^* := E_{H_0}(O_{ij}) = N \cdot \varphi_x(i) \cdot \varphi_y(j) \stackrel{?}{=} \frac{M_{i*}}{r} M_{*j}$  non si può calcolare!

↳ approssimo  $A_{ij}^*$  con la sua stima: devo stimare  $\varphi_x$  e  $\varphi_y$

$$\varphi_x(i) := P(X = i) \approx \frac{\#\{k : X_k = i\}}{N} = \frac{\sum_{j=1}^n O_{ij}}{N} = \frac{1}{N} M_{i*} \text{ marginale}$$

$$\varphi_y(j) \approx \frac{\sum_{i=1}^m O_{ij}}{N} = \frac{1}{N} M_{*j}$$

$$E_{H_0}(O_{ij}) \approx A_{ij}^* := N \cdot \frac{1}{N} M_{i*} \cdot \frac{1}{N} M_{*j} = \frac{1}{N} M_{i*} M_{*j} \quad \text{sono attesi "stimati"}$$

- la statistica è la solita  $W := \sum_{i,j} \frac{(O_{ij} - A_{ij})^2}{A_{ij}}$

- quanti gradi di libertà?

$$g.d.l = \# \text{categorie} - 1 - \# \text{parametri stimati}$$

$$= m \cdot n - 1 - (m-1) - (n-1) = mn - m - n + 1 = (m-1)(n-1)$$

righe e colonne meno 1 perché non sono parametri liberi: la somma fa 1

## ■ VERSIONI ESATTE DEI TEST DEL CHI-QUADRO

- al giorno d'oggi abbiamo strumenti per costruire delle versioni esatte (o quasi)
- in alcuni casi si fanno di combinatoria o per esaurizione dei casi
- in altri, l'unica possibilità è la simulazione Monte Carlo

- Caso base: esaurizione o MCS

esempio:  $X$ : titolo di studio

$$n = 20$$

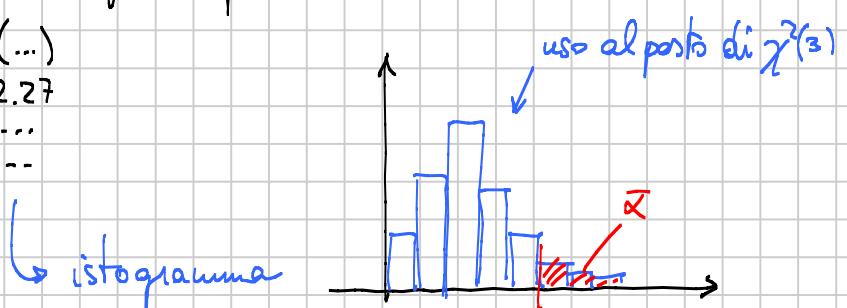
$\varphi_0$ :	medie	10%	1
	maturità	50%	2
	LT	30%	3
	LM	10%	4

$$A_1 = 10\% \cdot 20 = 2 \quad A_2 = 10 \quad A_3 = 6 \quad A_4 = 2 \quad \leftarrow \text{non vale la "rule of thumb"}$$

$$W = \sum_{i=1}^4 \frac{(O_i - A_i)^2}{A_i} = \frac{(O_1 - 2)^2}{2} + \frac{(O_2 - 10)^2}{10} + \frac{(O_3 - 6)^2}{6} + \frac{(O_4 - 2)^2}{2} = f(O_1, O_2, O_3, O_4)$$

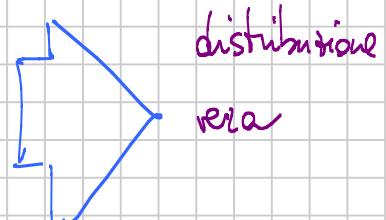
non mi fido di  $W \stackrel{H_0}{\sim} \chi^2(3)$  voglio sapere che distribuzione ha  $W$  sotto  $H_0$ :

a) MCS	id	$O_1 \ O_2 \ O_3 \ O_4$	$f(\dots)$
	1	2 5 9 4	2.27
	2	3 6 11 0	...
	:	.. .. ..	..
	M	↑	
		multin(20, $\varphi_0$ )	



b) esaurizione  $\text{id } O_1 \ O_2 \ O_3 \ O_4 \ P(\text{multin}(20, } \varphi_0) = 0 \text{ ) } f(\dots)$

1	0	0	0	20	$10^{-10}$
:	0	0	1	19	$10^{-9}$
:	..	..	..	..	
K	20	0	0	0	



## TEST ESATTO DI FISHER - IRWIN

nel caso particolare delle tabelle 2x2 si puo' fare il test in modo esatto

→ Supponiamo vera  $H_0$ , quindi  $X, Y$  indipendenti;  $n$  esperimenti;

Y: gruppo		m	f	
X: funzione	0	8 ⁵	8 ¹¹	16
	1	2 ⁵	4 ¹	6
		10	12	22

X		1	2	
Y	1	$O_{11}$	$O_{12}$	a
	2	$O_{21}$	$O_{22}$	b
		c	d	n

$$O_{22} = d - O_{12} = d - a + O_{11}$$

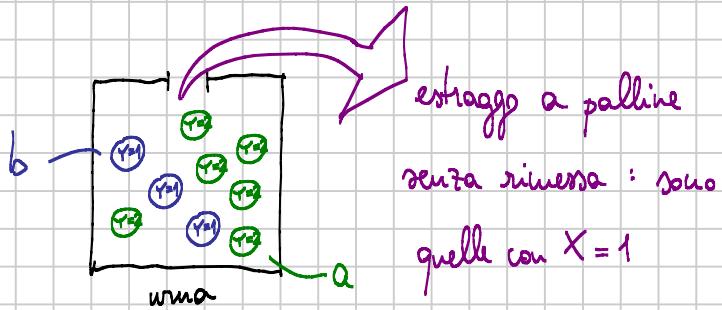
$$W = \frac{(O_{11} - \frac{ac}{n})^2}{\frac{ac}{n}} + \frac{(a - O_{11} - \frac{a \cdot d}{n})^2}{\frac{ad}{n}} + \frac{(c - O_{11} - \frac{bc}{n})^2}{\frac{bc}{n}} + \frac{(d - a + O_{11} - \frac{bd}{n})^2}{\frac{bd}{n}} = f(O_{11})$$

- per questo si usa  $O_{11}$  come statistica, invece di  $W$

- cerchiamo sotto  $H_0$  la distribuzione di  $O_{11}$

condizioniamo ai marginali

Y: gruppo		m	f	
X: funzione	0			16
	1			6
		10	12	22



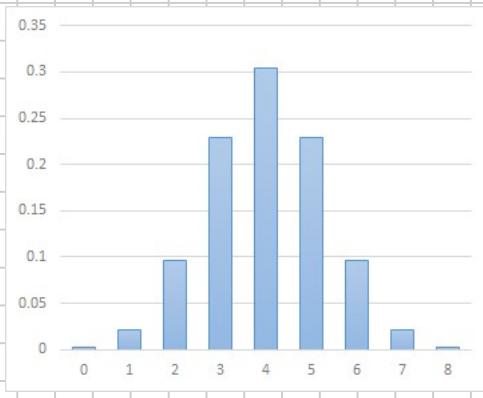
$$P_{H_0, \text{marg.}}(O_{11} = k) = \frac{\binom{a}{k} \binom{b}{c-k}}{\binom{n}{c}}$$

legge ipergeometrica

Ad esempio:  $a = 21$   $b = 21$   $c = 8$   $d = 34$   $n = 42$

valori osservati				
		4	17	21
		4	17	21
				21
osservati				8 34 42

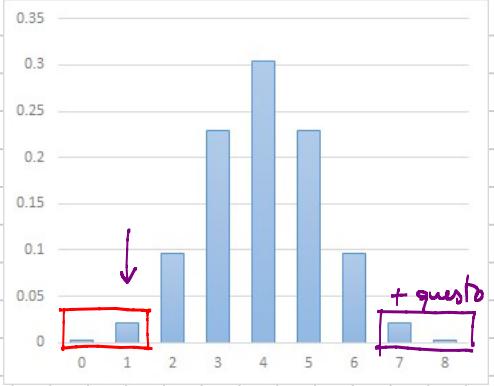
rule of thumb fallisce!



→ Supponiamo  $O_n = 1$  e calcoliamo il p-value

$$\alpha^* = P(O_n \leq 1) \cdot 2 \quad \text{perché è simmetrico}$$

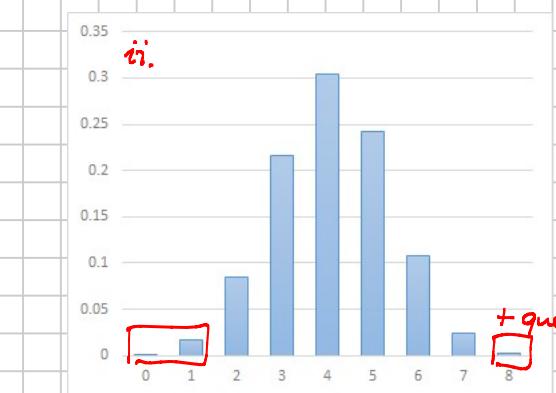
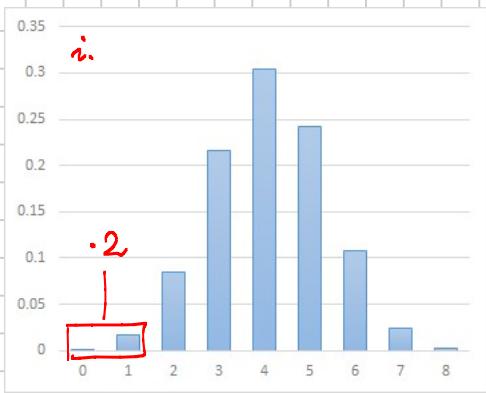
$\approx 4.48\%$



→ Quando non c'è simmetria, due possibilità

i.  $\alpha^* = 2 \cdot P(\text{coda minore})$

ii.  $\alpha^* = P(\text{coda minore}) + P(\text{altri esiti con prob. minore di quello osservato})$



# INTRODUZIONE AL MACHINE LEARNING

Note Title

ora 31

20/05/2025

## ① Visione d'insieme sul supervised learning

- dataset con input  $x$  e output  $y$
  - l'output si considera random  $y \rightarrow Y$ 
    - ↳  $Y$  con legge discreta : classification (legge multinomial)
    - ↳  $Y$  con legge continua : regression (legge Gaussiana)
    - * per leggi continue diverse, si assume di trasformare  $Y$  in Gaussiana
  - cose fatte :
    - regression
    - ANOVA
    - regressione logistica
 } regression classification
  - in tutti i casi si minimizza una funzione loss (-log-likelihood)
    - MSE vs cross-entropy
  - anche la differenza tra regressione e ANOVA (a 1 via, a 2 vie senza repl.) è minima e solo apparente :
- considero  $x$  categorica, con  $m$  categorie

$x$	$z_1$	$z_2$	$z_3$	$z_4$	$y$
4			1		~
1	1				~
3		1			~
1	1				~
⋮					~
2	1				~

$$\text{regressione} : Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \epsilon$$

$$\begin{aligned} \text{ANOVA} &: Y = \mu_x + \epsilon = \mu + \alpha_x + \epsilon \\ &= \mu + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \alpha_4 z_4 + \epsilon \end{aligned}$$

$$\text{differenza} : \beta_0 = 0 \quad \text{vs} \quad \sum_i \alpha_i = 0$$

HW: trovare le formule per convertire  $(\beta_0, \beta_1, \dots, \beta_{m-1})$  in  $(\mu, \alpha_1, \dots, \alpha_m)$

- Tutto insieme :

$$\begin{cases} z = b + w \cdot x \\ \lambda(b, w) = \text{loss}(y, z) \end{cases} \quad w, x \in \mathbb{R}^m \quad b, z, y \in \mathbb{R}$$

$\uparrow$  da minimizzare

z: stima media o logits

$$\text{loss}(y, z) = \begin{cases} (y - z)^2 \\ -y \log z - (1-y) \log(1-z) \end{cases}$$

MSE  
Binary C-E

- Anche vettoriale :

$$\begin{cases} z = b + w \cdot x \\ \lambda(b, w) = \text{loss}(y, z) \end{cases} \quad x \in \mathbb{R}^m \quad b, z, y \in \mathbb{R}^k \quad w \in M_{k,m}$$

$$\text{loss}(y, z) = \begin{cases} |y - z|^2 = \sum_{i=1}^k (y_i - z_i)^2 & \text{MSE multi-dim} \\ - \sum_{i=1}^k y_i \log z_i & \text{C-E} \end{cases}$$

- In tutti i casi si può considerare : regolarizzazione , validatione , ottimizzazione iterativa , ...

* Modello finale :

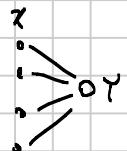
$$z = b + w \cdot x$$

ora 32

## ■ PRIMER SUL DEEP LEARNING

<https://www.skynettoday.com/overviews/neural-net-history>

'50-'60 singolo percezione  $Y = \Pi(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p > 0)$



'80 più layer si addestra con la backpropagation

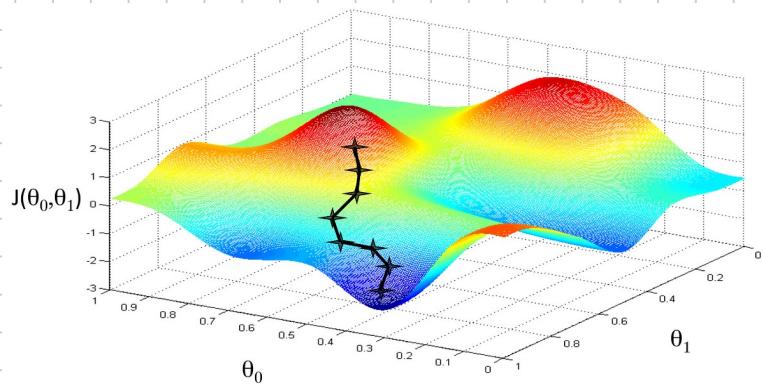
minimizzare la loss ?  $\rightarrow$  gradient descent

$$l(\beta_0 - \beta_{10}) \mapsto \nabla l$$

$$\frac{\partial l}{\partial \beta_i}$$

$$l = \frac{1}{N} \sum_{i=1}^N \left( y_i - \varphi(\beta_0 + \sum \beta_j \varphi(\beta_{j+1} + \sum \beta_k x_k)) \right)^2$$

backpropagation : algoritmo per calcolare  $\nabla l$

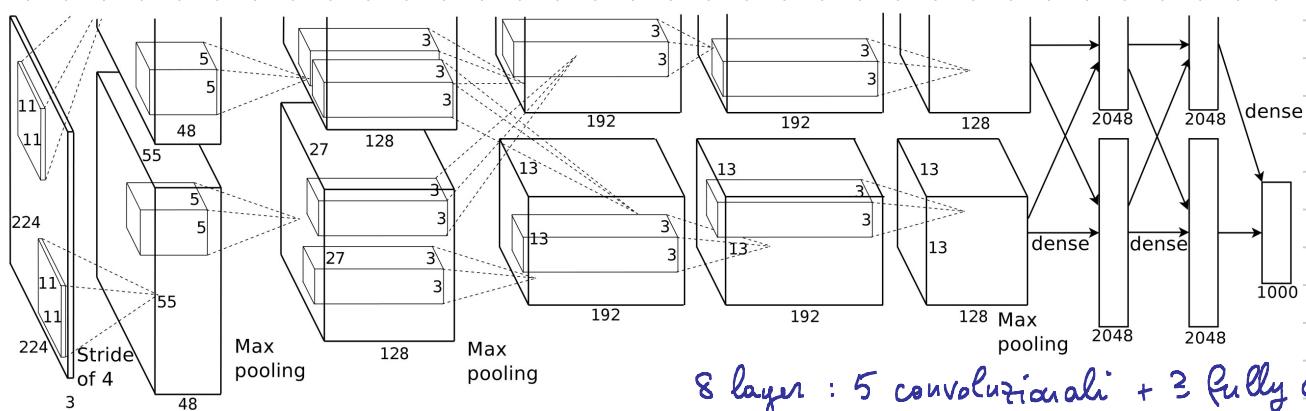
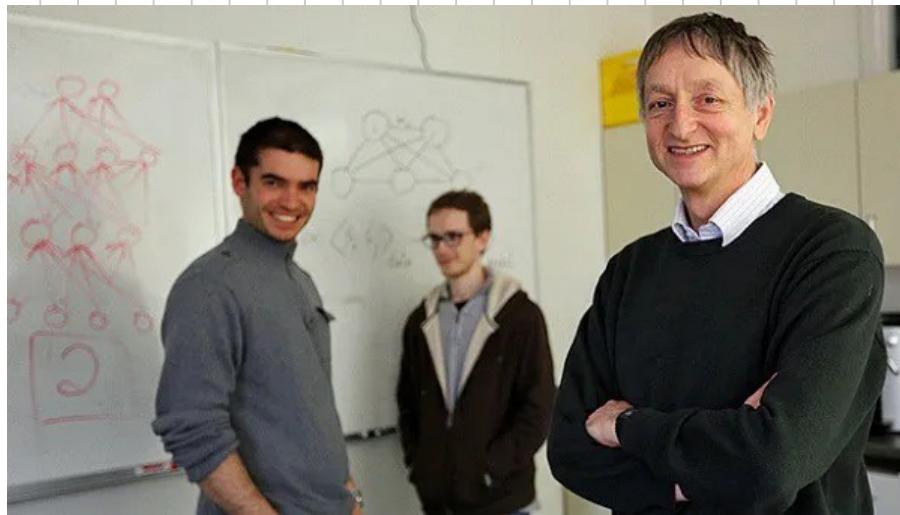


## ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky  
University of Toronto  
kriz@cs.utoronto.ca      Ilya Sutskever  
University of Toronto  
ilya@cs.utoronto.ca      Geoffrey E. Hinton  
University of Toronto  
hinton@cs.utoronto.ca

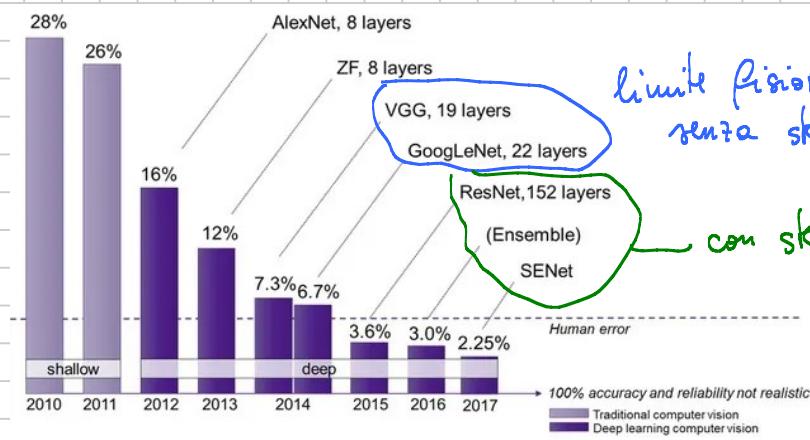
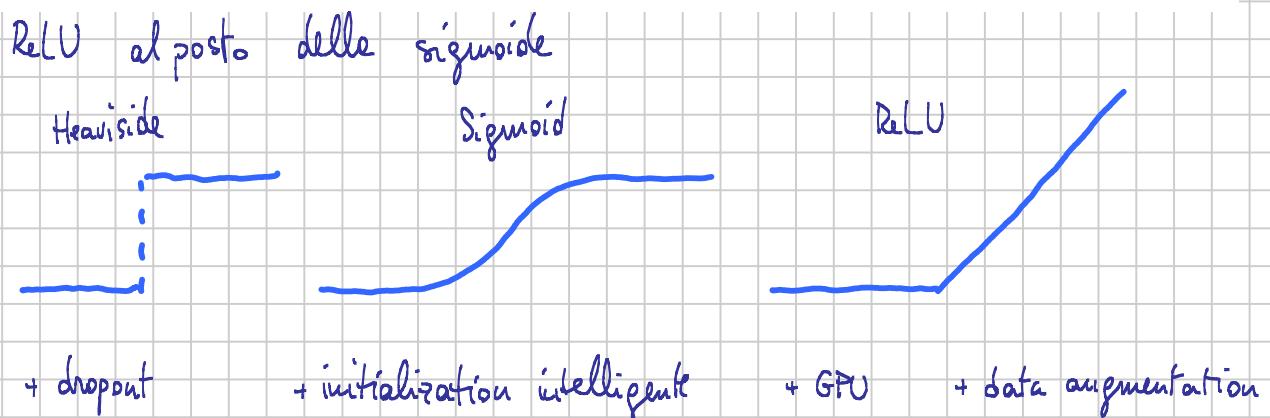
### Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called “dropout” that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.



8 layer : 5 convolutionali + 3 fully connected

Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network’s input is 150,528-dimensional, and the number of neurons in the network’s remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

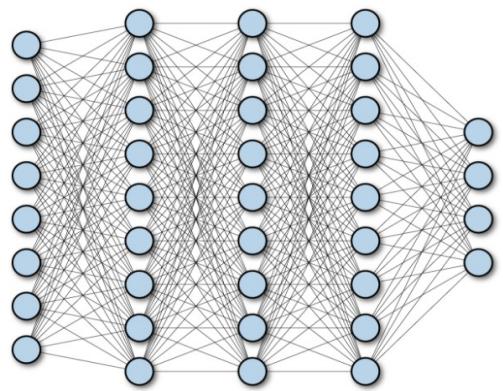
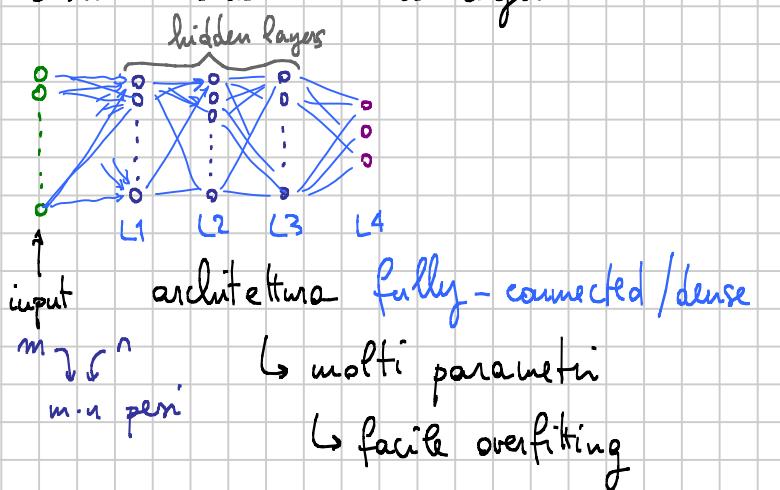


limite fisologico  
senza skip connection

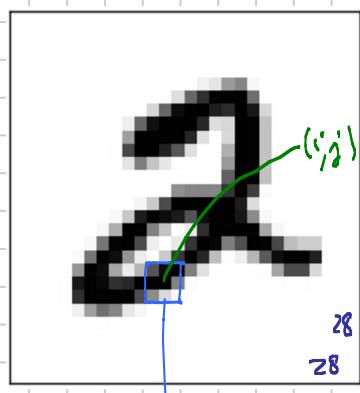
con skip connection

- <https://www.deeplearningbook.org/>
- <https://cs231n.github.io/convolutional-networks/>
- [https://github.com/vdumoulin/conv_arithmetic](https://github.com/vdumoulin/conv_arithmetic)
- <http://neuralnetworksanddeeplearning.com/>

## • Struttura canonica : a layer



## • Layer convoluzionale



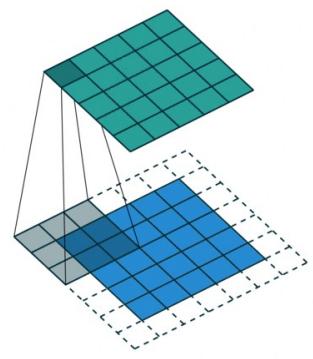
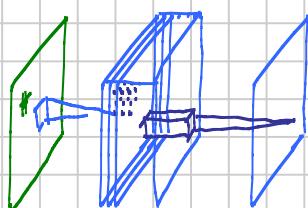
supponiamo : input fatto di variabili omologhe (ad esempio pixel 0-255) con struttura topologica



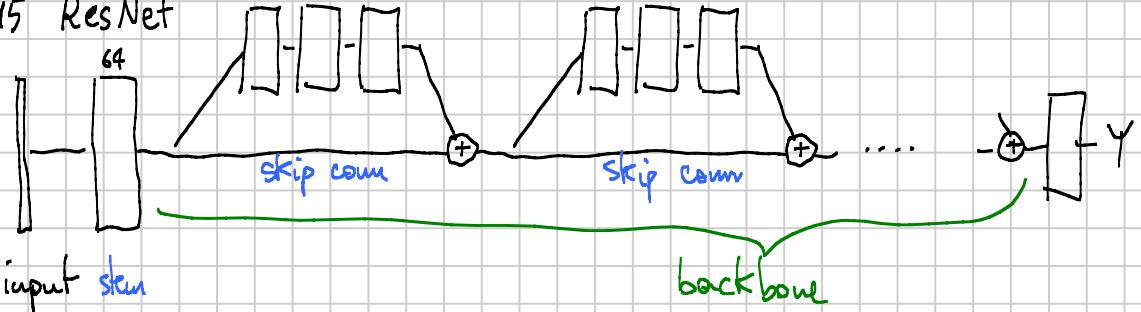
$$28^2 = 784 \text{ m}$$

$$z_{i,j}^{(l)} = \sum_{a=-1}^1 \sum_{b=-1}^1 w_{a,b} x_{i+a, j+b} \quad \forall i, j \quad \text{unità} : n = m$$

parametri : 9  
 weight sharing



2015 ResNet



Recenti : transformer

# Attention Is All You Need

Ashish Vaswani*  
Google Brain  
avaswani@google.com

Noam Shazeer*  
Google Brain  
noam@google.com

Niki Parmar*  
Google Research  
nikip@google.com

Jakob Uszkoreit*  
Google Research  
usz@google.com

Llion Jones*  
Google Research  
llion@google.com

Aidan N. Gomez* †  
University of Toronto  
aidan@cs.toronto.edu

Lukasz Kaiser*  
Google Brain  
lukaszkaiser@google.com

Illia Polosukhin* ‡  
illia.polosukhin@gmail.com

## Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

Alla base delle LLM  
moderne

Graph neural networks : Alpha fold 3