Analisi statistica

Riassunto delle slides

Contents

1 Calcolo numerico

1.1 Compilare un programma

Per eseguire un programma, un computer fa questa cosa: legge le informazioni dall'input e le mette nella memoria; poi da qui legge le istruzioni sequenzialmente e se c'è qualche calcolo da fare, lo fa fare alla ALU (aritmetic logic unit) e memorizza il risultato nella memoria; alla fine passa tutto in output.
Per essere eseguibile, un programma deve essere scritto in codice macchina e per questo si usano i linguaggi di programmazione che, attraverso il compilatore, vengono tradotti in codice macchina:

- codice sorgente,
- compilatore,
- codice oggetto,
- linker (aggiunge le librerie),
- codice eseguibile.

1.2 Errori di implementazione

Quando si passa dal modello astratto a quello implementato, si distinguono tre tipi di errori:

- errori analitici: causati dalla necessità di una aritmetica discreta, cioè il fatto che non si possa rappresentare una cosa continua e quindi la si fa a punti;
- errori inerenti o algoritmici (o di round-off): causati dal numero finito di cifre significative. Dipendono dall'algoritmo utilizzato, perché ci sono metodi migliori di altri per evitare questo tipo di problemi.

1.3 Rappresentazione di un numero

L'obiettivo è quello di tenere sotto controllo gli errori di round-off. Per rappresentare un numero, esistono diverse rappresentazioni, tra cui quella di Von Neumann: si dedicano n cifre significative alla mantissa e poi un tot anche alla caratteristica, cioè all'ordine di grandezza (che generalmente è in base 2).
Con n cifre significative in base 2, il numero più alto rappresentabile è $2^n - 1$ perché con due cifre (0 e 1), ci sono 2^n possibili numeri rappresentabili e, se posti in ordine crescente, ognuno dista dal precedente un'unità, con il minimo che vale 0 (tutti 0), per cui il numero più alto che si può rappresentare è $2^n - 1$:

- 0 è l'1,
- 1 e il 2,
- ...$2^n - 1$ è il 2^n-esimo.

questo numero corrisponde a un numero in base dieci di m cifre significative, per cui:

$$m = \log(2^n - 1)$$

quindi per un numero con 24 bit per la mantissa, si hanno circa 7 cifre significative in base 10.
In precisione semplice (floating point):

- un bit per il segno,
- 23 bit per la mantissa (quindi 7 cifre significative),
- 8 bits per la caratteristica.

In precisione doppia (duble) si raddoppiano i bite alla mantissa e le cifre significative diventano 17, però il tempo per le moltiplicazioni è triplicato e si occupa più memoria, quindi non è che convenga moltissimo.

Esiste anche la rappresentazione fixed-point, in cui c'è un bit per il segno e i restanti per il numero in sé, in cui da qualche parte c'è la virgola. Nel caso del fixed pointd, l'errore relativo può variare moltissimo a seconda del numero rappresentato, mentre per i float è lo stesso per ogni numero: la densità relativa non è costante.
Quando un insieme di rappresentazione non è chiuso rispetto ad un'operazione, il risultato presenta sicuramente errore di round-off e deve essere approssimato.

1.4 Errori algoritmici

Per numeri troppo grandi o troppo piccoli che non si riesce a rappresentare, si parla rispettivamente di overflow o underflow. Per questo motivo, somme e soprattutto differenze di numeri grandi portano grandi errori dovuti all'arrotondamento: bisogna cercare di evitarlo se possibile. Per esempio, per calcolare un polinomio, conviene usare il metodi di Ruffini-Horner perché riduce il numero di operazioni da effettuare:

- metodo semplice:

  $$p(x) = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n$$

  che sono n addizioni e n(n+1)/2 moltiplicazioni, quindi $\sim n^2$.

- metodo di R-H:

  $$p(x) = (((((a_n)x + a_{n-1})x + a_{n-2}x + ...$$

  che sono n addizioni e n moltiplicazioni, quindi $\sim n$.

Il problema si verifica anche se si sottraggono due numeri molto simili tra loro, perché il risultato è un numero molto piccolo che potrebbe finire in underflow (si parla di "cancellazione catastrofica").

oss: Il costo computazionale della risoluzione di un sistema lineare in n equazioni è asintoticamente uguale al costo del prodotto di due matrici nxn. Esistono algoritmi

che non richiedono più di $k \cdot n^{\alpha}$ operazioni, col il più piccolo valore noto di α pari a 0.2375.

Alcuni algoritmi sono definiti "instabili": accade quando gli errori di round-off si accumulano fino a portare a risultati completamente errati.
Esistono problemi detti "mal condizionati" che con qualsiasi algoritmo danno errori talmente elevati da rendere il risultato privo di significato. In questi casi, piccole variazioni dei dati iniziali portano a grandi variazioni nei risultati. Si chiama "numero di condizionamento del problema" il seguente rapporto:

$$\frac{\Delta r}{\Delta \alpha} = \frac{\%\text{ errore risultato}}{\%\text{ errore dato iniziale}}$$

che può quindi essere espresso in questo modo:

$$\Delta\alpha = \frac{x+h-x}{x} = \frac{h}{x} \qquad \Delta r = \frac{f(x+h)-f(x)}{f(x)}$$

$$\frac{\Delta r}{\Delta \alpha} = \frac{f(x+h)-f(x)}{h} \cdot \frac{x}{f(x)} = f'(x) \cdot \frac{x}{f(x)}$$

Se questo numero è $<< 1$, allora il problema è poco sensibile ai dati iniziali.

2 Probabilità

2.1 Concetto di probabilità

Consideriamo un insieme S di eventi detto "spazio campionario" e due eventi casuali A e B. Essi sono soggetti agli assiomi di Kolmogorov:

$\forall A \subseteq S, 0 \leqslant P(A) \leqslant 1$
$P(S) = 1$
$A \cap B = \emptyset \Longrightarrow P(A \cup B) = P(A) + P(B)$

Si definisce probabilità condizionata di A dato B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Se due eventi sono indipendenti, allora vale che:

$$P(A \cap B) = P(A) \cdot P(B) \Longrightarrow P(A|B) = \frac{P(A) \cdot P(B)}{P(B)} = P(A)$$

Esistono diversi approcci per definire la probabilità:

- approccio frequentista:

  $$P(A) = \lim_{n \to +\infty} \frac{\#A}{n}$$

  ma c'è il problema che è impossibile fare un numero infinito di tentativi.

- approccio bayesiano:
  utilizzare il teorema di Bayes, per esempio:

  $$P(\text{teoria}|\text{dati}) \propto P(\text{dati}|\text{teoria}) \cdot P(\text{teoria})$$

  si tratta quindi di un approccio "soggettivo", almeno per l'ultimo termine dell'equazione qua sopra, perché la probabilità della validità della teoria può essere data sulla base di ragionamenti e osservazioni riguardanti il fenomeno di cui si sta parlando.

2.1.1 Esempio di probabilità condizionata

Si osservano dei decadimenti e se ne misurano indipendentemente N_a e N_b; in comune ne sono stati misurati N_{ab}. Possiamo stimare il numero totale di eventi N e l'efficienza totale ϵ.

$$P(a) = \frac{N_a}{N} \qquad\qquad P(b) = \frac{N_b}{N}$$

$$P(ab) = \frac{N_{ab}}{N} = P(a) \cdot P(b) = \frac{N_a \cdot N_b}{N^2} \implies N = \frac{N_a \cdot N_b}{N_{ab}}$$

$$\epsilon = P(a \cup b) = P(a) + P(b) - P(a \cap b) = \frac{N_a + N_b - N_{ab}}{N}$$

2.2 Teorema di Bayes

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad\qquad P(B|A) = \frac{P(B \cap A)}{P(A)}$$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Dal teorema di Bayes è possibile dedurre il teorema della probabilità totale: consideriamo un insieme S diviso in sottoinsiemi disgiunti A_i la cui unione dà l'insieme di partenza:

$$\cup_i A_i = S$$

e consideriamo un insieme B anch'esso interno ad S:

$$B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i)$$
$$\implies P(B) = \sum_i P(B \cap A_i)$$
$$\implies P(B) = \sum_i P(B|A_i) P(A_i)$$

da cui, per uno specifico insieme A_j, attraverso il teorema di Bayes:

$$P(A_j|B) = \frac{P(B|A_j) \cdot P(A_j)}{\sum_i P(B|A_i) P(A_i)}$$

2.2.1 Esempio di probabilità totale: segnale e fondo

Un rilevatore misura segnale e fondo con relative efficienze $P(R|S)$ e $P(R|F)$. Se sono note a priori la probabilità di segnale e di fondo $P(S)$ e $P(F)$, allora si può risalire alla probabilità, data una misurazione, di aver misurato il segnale:

$$P(S|R) = \frac{P(R|S) \cdot P(S)}{P(R|S) \cdot P(S) + P(R|F) \cdot P(F)}$$

2.2.2 Esempio di probabilità totale: AIDS

$$P(\text{AIDS}) = 0.001 \qquad P(\text{no AIDS}) = 0.999$$
$$P(+|\text{AIDS}) = 0.98 \qquad P(\text{-}|\text{AIDS}) = 0.02$$
$$P(+|\text{no AIDS}) = 0.03 \qquad P(\text{-}|\text{no AIDS}) = 0.97$$

Quindi se il test è positivo, la probabilità di avere davvero preso l'AIDS è:

$$P(\text{AIDS}|+) = \frac{P(+|\text{AIDS}) \cdot P(\text{AIDS})}{P(+|\text{AIDS}) \cdot P(\text{AIDS}) + P(+|\text{no AIDS}) \cdot P(\text{no AIDS})} = 0.032$$

2.2.3 Approccio bayesiano con una teoria

Cominciamo com un esempio: con un esperimento è stata misurata la massa di un elettrone e sono stati trovati i valori $\{m_i\}$, di cui il valore medio è $m_e = 520(10)\,\text{KeV}$. Si assume quindi che il valore vero sia compreso tra 510 KeV e 530 KeV con una probabilità del 68% data dal confidence level. Per cui, la probabilità che la massa vera sia proprio 520 KeV è data da:

$$P(m_e|m_i) =$$

rivedere negli appunti questa formula

Quando si fa una misurazione, si misurano N valori e si calcolano il valore medio $\bar{x}$ e la deviazione standard σ e si dice che il risultato è:

$$\mu = \bar{x} \pm \frac{\sigma}{\sqrt{N}}$$

e di solito lo si interpreta dicendo che:

$$P\left(\bar{x} - \frac{\sigma}{\sqrt{N}} \leqslant \mu \leqslant \bar{x} + \frac{\sigma}{\sqrt{N}}\right) = 68\%$$

ma in realtà quello che sappiamo è solo che:

$$P\left(\mu - \frac{\sigma_{true}}{\sqrt{N}} \leqslant \bar{x} \leqslant \mu + \frac{\sigma_{true}}{\sqrt{N}}\right) = 68\%$$

WHAT?!?

3 Statistica

3.1 Distribuzioni di probabilità

Una funzione di densità di probabilità f è definita in modo che la probabilità che una variabile x sia compresa tra x e $x + dx$ sia data da:

$$P(x \subset [x, x + dx]) = f(x)dx$$

dunque vale che:

$$\int_{-\infty}^{+\infty} dx f(x) = 1$$

Si definisce funzione cumulante:

$$F(x) = \int_{-\infty}^{x} dx' f(x')$$

e quantile di ordine α il valore di x per cui $F(x) = \alpha$.
Nel caso multidimensionale in cui si abbiano due o più variabili, si parla di joint pdf:

$$f(x, y) \qquad \Longrightarrow \qquad \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} dxdyf(x, y) = 1$$

e si definiscono due distribuzioni marginali:

$$f_x(x) = \int_{-\infty}^{+\infty} dy f(x, y) \qquad\qquad f_y(y) = \int_{-\infty}^{+\infty} dx f(y, x)$$

dunque due variabili x e y sono indipendenti se $f(x) = f_x(x) \cdot f_y(y)$. Ora, se A è l'evento di probabilità $f_x(x)dx$, mentre B ha probabilità $f_y(y)dy$, allora si possono definire le pdf condizionali come segue:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{f(x, y)dxdy}{f_x(x)dx} \qquad \Longrightarrow \qquad h(y|x) = \frac{f(x, y)}{f_x(x)}$$

per cui il teorema di Bayes diventa:

$$g(x|y) = \frac{h(y|x)f_x(x)}{f_y(y)}$$

3.1.1 Paradosso di Borel-Kolmogorov

Si considerino dei punti distribuiti uniformemente sulla superficie del pianeta Terra: ci si aspetterebbe che i punti siano uniformemente distribuiti anche lungo un parallelo o un meridiano... ma consideriamo un meridiano: esso giace per il 25% a nord del 45'esimo parallelo e quindi, secondo la logica di prima, anche il 25% dei punti che si trovano su di esso. Però non è vero che il 45% della superficie terrestre è al di sopra del 45'esimo parallelo!
Il paradosso è risolto perché non ci si può basare su un insieme di misura nulla quale il meridiano (perché è unidimensionale). Lo si vede chiaramente adottando la terminologia poc'anzi introdotta:
Se la distribuzione è uniforme, la probabilità di trovare un punto in una certa superficie è dato dal rapporto tra l'angolo solido descritto da tale superficie e l'angolo solido totale:

$$f(\theta,\phi)d\theta d\phi = \frac{d\phi d\theta\cos(\theta)}{4\pi}$$

da cui è possibile determinare la due probabilità marginali:

$$f_\phi(\phi) = \int_0^\pi f(\theta,\phi)d\theta = \int_0^\pi \frac{\cos(\theta)}{4\pi} = \frac{\cos(\theta)}{2}$$

$$f_\theta(\theta) = \int_0^{2\pi} f(\theta,\phi)d\phi = \frac{1}{2\pi}$$

per cui si tratta di due costanti rispetto alle rispettive variabili. Da ciò si può dunque dedurre che, mentre la densità lungo un parallelo è effettivamente costante, lo stesso non si può dire riguardo a un meridiano.

Una funzione di una variabile casuale è essa stessa una variabile casuale. Consideriamo la pdf $f(x)$ e una funzione $a(x)$ di cui si vuole trovare la pdf $g(a)$. Nel caso in cui l'inversa di $a(x)$ sia univoca, definita dS la regione delle x per cui $a \subset [a, a+da]$:

$$g(a)da = \int_{dS} dx f(x) = \left| \int_{x(a)}^{x(a+da)} f(x')dx' \right| = \int_{x(a)+|\frac{dx}{da}|da}^{x(a+da)} f(x')dx'$$

Ovvero:

$$g(a) = f(x(a)) \left| \frac{dx}{da} \right|$$

e se $x(a)$ non è univoca, allora bisogna considerare tutti gli intervalli dS di dx che corrispondono a da.
Nel caso di funzioni di N variabili, siccome vale che:

$$g(a')da' = \int ... \int_{dS} f(x_1 ... x_N) dx_1 ... dx_N$$

con dS regione dello spazio delle x compreso tra le isosuperfici:

$$a(\vec{x}) = a' \quad \wedge \quad a(\vec{x}) = a' + da'$$

Nel caso in cui $z = x \cdot y$, si trova la convoluzione di Mellin:

$$g(z)dz = \int_{dS} dxdy f(x,y) = \int_{-\infty}^{+\infty} dx \int_{\frac{z}{x}}^{\frac{z+dz}{x}} dy f(x,y)$$

Non ho capito questa parte...

3.2 Propagazione degli errori

Consideriamo una variabile x con pdf $f(x)$. Si definisce valore di aspettazione o media (e lo si indica spesso con μ):

$$E[x] = \int dx f(x) x$$

Nel caso di una variabile $y(x)$ con pdf $g(x)$, invece:

$$E[y] = \int dy \cdot y \cdot g(y) = \int dx f(x) g(x)$$

Mentre si definisce varianza (e la si indica spesso con σ^2, mentre con deviazione standard si intende σ):

$$V[x] = E[x - E[x]^2] = E[x^2] - \mu^2$$

Più in generale si definiscono 'momenti algebrici' $E[x^n] = \mu'_n$ con $\mu'_1 = \mu$ e 'momenti centrali' $E[(x-\mu)^n] = \mu_n$ con $\mu_2 = \sigma^2$.
Si definiscono inoltre due grandezze di correlazione. La covarianza:

$$\mathrm{cov}[x, y] = E[xy] - E[x]E[y] = E[xy] - \mu_x \mu_y$$

che equivale a:

$$\begin{aligned}
\mathrm{cov}[x, y] &= E[(x - \mu_x)(y - \mu_y)] \\
&= E[xy - x\mu_y - y\mu_x + \mu_x \mu_Y] \\
&= E[xy] - \mu_y E[x] - \mu_x E[y] + \mu_x \mu_y \\
&= E[xy] - \mu_y mu_x - \mu_x \mu_y + \mu_x \mu_y \\
&= E[xy] - \mu_x \mu_y
\end{aligned}$$

Notare che se x e y sono indipendenti, allora $f(x, y) = f_x(x) f_y(y)$, perciò:

$$E[xy] = \int dx \int dy xy f(x, y) = \mu_x \mu_y \qquad \Longrightarrow \qquad \mathrm{cov}[x, y] = 0$$

e il coefficiente di correlazione:

$$\rho_{xy} = \frac{\mathrm{cov}[xy]}{\sigma_x \sigma_y}$$

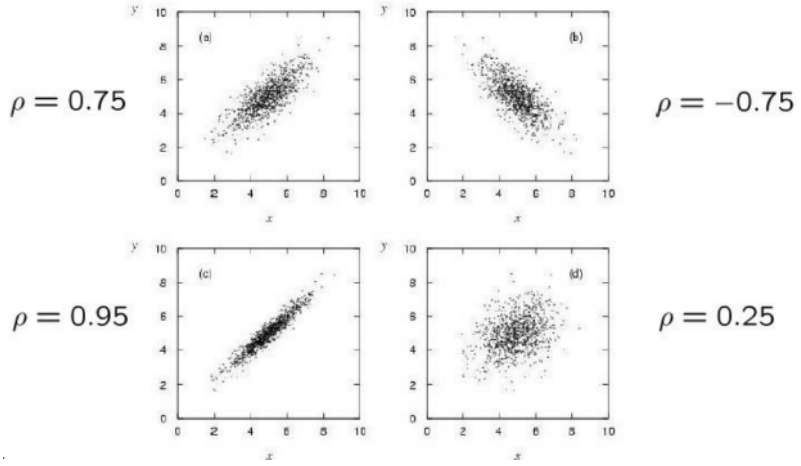


Figure 1: Esempio di correlazione tra due grandezze.

Anche se la $f(\vec{x})$ non è completamente nota, è comunque possibile stimare il valore medio e la varianza di una grandezza $y(\vec{x})$ conoscendo solo le stime di media e varianza della pdf. Espandiamo attraverso la serie di Taylor:

$$y(\vec{x}) = y(\vec{\mu}) + \sum_{i=1}^{N} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i)$$
$$\implies \qquad E[y] = y(\vec{\mu}) \impliedby \qquad E[x_i] = \mu_i$$

Mentre per la varianza servono $E[y^2]$ ed $E[y]$. Sempre passando attraverso uno sviluppo di Taylor attorno al valore medio:

$$E[y^2] = y^2(\vec{\mu}) + 2y(\vec{\mu}) \sum_{i=1}^{N} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} E[x_i - \mu_i]$$
$$+ E\left[\left(\sum_{i_1}^{N} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i)\right)\left(\sum_{j=1}^{N} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} (x_j - \mu_j)\right)\right]$$

Siccome il secondo termine si annulla sempre perché $E[x_i] = \mu_i$, allora rimane che:

$$V[y] = E[y^2] - E[y]^2 = \sigma_y^2 = \sum_{i,j=1}^{N} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} V_{ij}$$

Con V_{ij} che è la matrice di covarianza, che ha come entrate:

$$V_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)] = \rho_{ij}\sigma_i\sigma_j$$

e quindi, nel caso in cui le variabili siano scorrelate, si ottiene che:

$$V_{ij} = \sigma_i^2 \delta_{ij} \qquad \implies \qquad \sigma_y^2 = \sum_{i=1}^{N} \left[\frac{\partial y}{\partial x_i}\right]^2_{\vec{x}=\vec{\mu}} \sigma_i^2$$

Cioè dice quanto cambia la y al variare del 'dato iniziale' $\vec{x}$. Ma quindi, per quanto visto prima:

$$\mathrm{cov}[x_i, x_j] = E[(x_i - \mu_i)(x_j - \mu_j)] = V_{ij}$$

Più in generale, date $\vec{y}$ variabili dipendenti da $\vec{x}$, vale che:

$$U = AVA^T \qquad \text{con} \qquad A_{ij} = \left[\frac{\partial y_i}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} \qquad \text{e con} \qquad U_{kl} = \mathrm{cov}[y_k, y_l]$$

dove U è detta matrice di covarianza delle y.
Attenzione: quanto detto fin'ora, che descrive in che modo gli errori di $\vec{x}$ influenzano

y, vale solo nel caso in cui y sia lineare nelle x. Quindi, in casi come $y(x) = 1/x$, non si può fare questo discorso.

3.2.1 Esempio

Consideriamo:

$$y = x_1 - x_2$$
$$\text{con} \qquad \mu_1 = \mu_2 = 10 \qquad \wedge \qquad \sigma_1 = \sigma_2 = 1$$

allora abbiamo che $y = y(x_1, x_2)$, quindi:

$$E[y] = y(\mu_1, \mu_2) = 10 - 10 = 0$$
$$V[y] = \sum_{i,j=1}^{2} \left[\frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij} = 1 \cdot V_{11} + 1 \cdot V_{22} - 1 \cdot 2 \cdot V_{12}$$

Se le correlazioni sono nulle, allora $V_{12} = 0 \implies V[y] = 2 \implies \sigma_y = 1.4$, se invece x_1 e x_2 sono correlate, nel caso in cui il coefficiente di correlazione sia unitario si ha che $V[y] = 0$. Quindi la correlazione può cambiare di molto le cose.

3.2.2 Errori sistematici

Consideriamo due grandezze x_1 e x_2 con un errore sistematico in comune S:

$$x_1 = x_{1_0} + x_{1_s}$$
$$x_2 = x_{2_0} + x_{2_s}$$

si avrà che i termini con pedice 0 sono indipendenti tra loro, mentre gli altri due saranno correlati. Dato che gli errori si sommano in quadratura, la matrice di covarianza sarà quindi:

$$\text{cov}[x_1, x_2] = S^2 \qquad \implies \qquad V = \begin{pmatrix} \sigma_1^2 + S^2 & S^2 \\ S^2 & \sigma_2^2 + S^2 \end{pmatrix}$$

perché:

$$
\begin{aligned}
\mathrm{cov}[x_1, x_2] &= E[x_1 x_2] - E[x_1]E[x_2] = \\
&= E[(x_{1_0} + x_{1_s})(x_{2_0} + x_{2_s})] - E[x_{1_0} + x_{1_s}]E[x_{2_0} + x_{2_s}] = \\
&= E[x_{1_0} x_{2_0}] + E[x_{1_0} x_{2_s}] + E[x_{1_s} x_{2_0}] + E[x_{1_s} x_{2_s}] + \\
&\quad - E[x_{1_0}]E[x_{2_0}] - E[x_{1_0}]E[x_{2_2}] - E[x_{1_s}]E[x_{2_0}] - E[x_{1_s}]E[x_{2_s}] = \\
&= \mu_1\mu_2 + \mu_1 E[x_{2_s}] + E[x_{1_s}]\mu_2 + E[x_{1_s} x_{2_s}] + \\
&\quad - \mu_1\mu_2 - \mu_1 E[x_{2_s}] - E[x_{1_s}]\mu_2 - E[x_{1_s}]E[x_{2_s}] = \\
&= E[x_{1_s} x_{2_s}] - E[x_{1_s}]E[x_{2_s}] = \mathrm{cov}[x_{1_s}, x_{2_s}]
\end{aligned}
$$

3.2.3 Trasformazione ortogonale

Può tornare utile fare un cambio di variabile che permetta di ottenere una matrice di covarianza delle y diagonale.
Consideriamo le solite variabili Y_i legate linearmente alle x_j:

$$
y_i = \sum_j = A^i_j x_j \qquad \Longrightarrow \qquad U_{ij} = \sum_{k,l} A_{ik} V_{kl} A^T_{lj}
$$

Si tratta quindi di diagonalizzare la matrice U: la soluzione è semplice, la matrice A è quella formata dagli autovalori di V. Questo concetto è utile nel caso della scelta delle coordinate da utilizzare.
Se immaginiamo di star utilizzando le coordinate polari $\vec{x} = (x, y)$, la matrice di covarianza sarà:

$$
V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}
$$

Diagonaliziamola: prima di tutto troviamo gli autovalori della matrice $V - \lambda I$:

$$
\begin{vmatrix} \sigma_1^2 - \lambda & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 - \lambda \end{vmatrix} = (\sigma_1^2 - \lambda)(\sigma_2^2 - \lambda) - \rho^2\sigma_1^2\sigma_2^2 = 0
$$

$$
\begin{aligned}
\Longrightarrow \qquad & \lambda^2 - (\sigma_1^2 + \sigma_2^2)\lambda + \sigma_1^2\sigma_2^2(1 - \rho^2) = 0 \\
\Longrightarrow \qquad & \lambda_{1,2} = \frac{\sigma_1^2 + \sigma_2^2 \pm \sqrt{\sigma_1^4 + \sigma_2^4 + 2\sigma_1^2\sigma_2^2 - 4\sigma_1^2\sigma_2^2 + 4\rho^2\sigma_1^2\sigma_2^2}}{2} = \\
\Longrightarrow \qquad & \lambda_{1,2} = \frac{\sigma_1^2 + \sigma_2^2 \pm \sqrt{(\sigma_1^2 - \sigma_2^2)^2 + 4\rho^2\sigma_1^2\sigma_2^2}}{2}
\end{aligned}
$$

e ora calcoliamo gli autovettori:

$$
(V - \lambda I)\vec{r} = 0 \qquad \Longrightarrow \qquad \begin{pmatrix} \sigma_1^2 - \lambda & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 - \lambda \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}
$$

$$\implies \qquad \begin{cases} (\sigma_1^2 - \lambda) r_1 + \rho \sigma_1 \sigma_2 r_2 = r_1 \\ \rho \sigma_1 \sigma_2 r_1 + (\sigma_2^2 - \lambda) r_2 = r_2 \end{cases} \qquad \implies \qquad r_1 = r_2$$

eccetera eccetera...

Distribuzioni di probabilità

3.3 Distribuzione binomiale

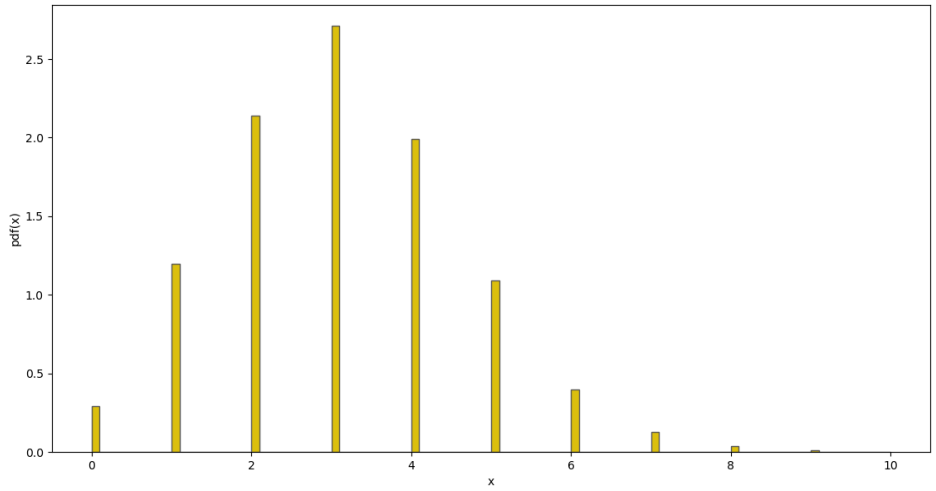


Figure 2: Distribuzione binomiale: p = 0.3, n = 10, N = 1000.

Si considerino N tentativi di un esperimento che può avere come esiti soltanto successo o fallimento e che la probabilità di ogni successo sia p. Definiamo n il numero dei successi. Dunque la probabilità di ottenere n successi su N tentativi totali è data da:

$$P(N, n, p) = \binom{N}{n} p^n (1-p)^{N-n}$$

perché:

- la probabilità che un successo si verifichi è p;
- la probabilità che k successi si verifichino è data dal prodotto di tutte le probabilità: p^n;
- lo stesso discorso vale per gli insuccessi: ognuno ha probabilità (1_p) e se ne verificano $N - n$;
- il termine binomiale rappresenta tutte le possibili permutazioni: il concetto è semplice se si immagina di posizionare successi e fallimenti all'interno di

una griglia con N possibili posizioni: un successo è un pallino bianco e un fallimento è un pallino nero. In quanti posti posso mettere il primo successo? N. E il secondo? $N-1$. E il terzo? E così via, finché ho messo tutti i successi, che occupano n posizioni, l'ultima delle quali è stata scelta tra $N-(n-1)$ posizioni, per cui:

$$N \cdot (N-1) \dots (N-n+1) = \frac{N \cdot (N-1) \dots (N-n+1) \cdot 2 \cdot 1}{(N-n) \cdot (N-n-1) \dots 2 \cdot 1} = \frac{N!}{(N-n)!}$$

ma non bisogna considerare che poi tutte le posizioni delle palline nere sono uguali, quindi va ulteriormente diviso per $n!$ per le stesse ragioni. Da cui:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

Per la normalizzazione, vale che:

$$\sum_{n=0}^{N} P(N,n,p) = 1$$

Possiamo definire un valore di aspettazione e una varianza:

$$E[n] = \sum_{n=0}^{N} nP(N,n,p) = Np$$
$$V[n] = E[n^2] - E[n]^2 = Np(1-p)$$

3.4 Distribuzione multinomiale

È la generalizzazione della pdf precedente nel caso in cui ci siano m possibili risultati, ciascuno con una probabilità P_m di verificarsi. Per esempio, è il caso di un istogramma riguardo al quale ci si domanda quale sia la probabilità di trovarlo esattamente con quelle specifiche entrate.
La probabilità è conseguentemente data da:

$$P(N,\vec{n},\vec{p}) = \frac{N!}{n_1!n_2!\dots n_m!} p_1^{n_1} p_2^{n_2} \dots p_m^{n_m}$$

E come valore di aspettazione e deviazione standard si ottiene che:

$$E[n_i] = Np_i$$
$$V[n_i] = Np_i(1-p_i)$$

3.4.1 Legge dei grandi numeri

La legge dei grandi numeri afferma che la media sperimentale di una variabile x, per un numero di tentativi N che tende all'infinito, si avvicina molto alla media vera. Questa legge può essere utilizzata per stimare le probabilità P_i di una distribuzione muiltinomiale tramite le frequenze con cui i diversi eventi si verificano.
Si consideri la frequenza f_j con cui l'evento j-esimo si verifica, dato un set di N tentativi:

$$f_j = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_j}{N}$$

dove x_i è una variabile che vale 1 se l'evento j-esimo si è verificato e vale 0 quando se ne è verificato un altro e x_j è quindi il numero di volte che l'evento j-esimo si è verificato.
A differenza di P_j, f_j è una variabile casuale perché dipende da x_j che è la somma di variabili casuali. Definiamo valore medio:

$$E(f_j) = \frac{E(x_j)}{N} = P_j$$

e calcoliamo la varianza:

$$V\left[f_j\right] = V\left[\frac{x_j}{N}\right] = E\left[\frac{x_j^2}{N^2}\right] - \left(E\left[\frac{x_j}{N}\right]\right)^2 = \frac{1}{N^2}V\left[x_j\right]$$

ora, x_j è esattamente n_j della multinomiale, perciò:

$$V[x_j] = NP_j(1-P_j) \qquad \Longrightarrow \qquad V[f_j] = \frac{1}{N}P_j(1-P_j) \leqslant \frac{1}{N}$$

3.5 Distribuzione di Poisson

Se si considera la distribuzione binomiale e ci si pone nel limite in cui il numero di tentativi ripetuti tenda all'infinito e che la probabilità di successo tenda a zero (con il vincolo che $N \cdot p = cost = \nu$), si ottiene la distribuzione di Poisson:

$$P(N, n, \nu) = \frac{\nu^n}{n!}e^{-\nu}$$

con:

$$E[n] = \nu$$
$$V[n] = \nu$$

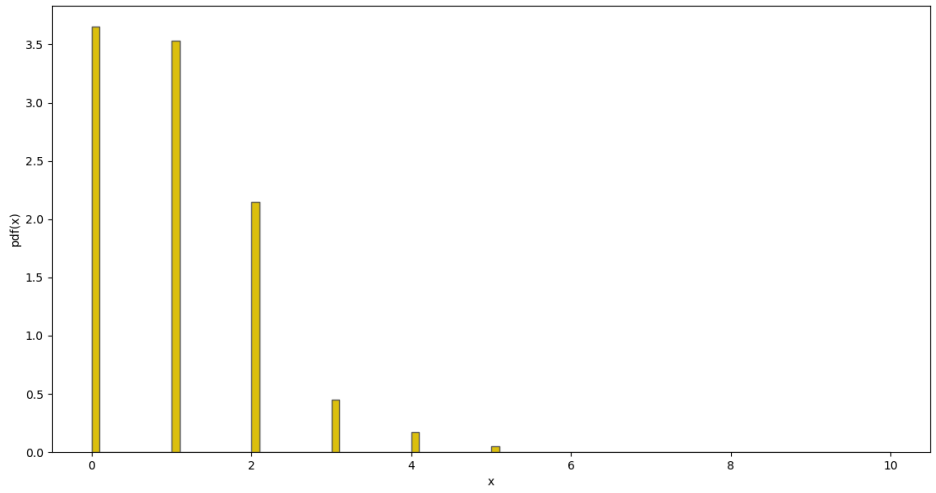


Figure 3: Distribuzione Poissoniana: $\mu = 1$, N = 1000.

dove $\nu = NP =$ numero medio di successi.
Quando N è talmente grande da non essere definito (come nel caso in cui si osservino i decadimenti di un atomo e dunque i tentativi sono le osservazioni, che sono dunque continue), non è più possibile definire una probabilità di successo per ogni evento (perché sarebbe nulla, da cui il motivo per cui la Poissoniana è definita con questi due limiti), e quindi ν va definita in un altro modo. Infatti la distribuzione di Poisson, trattandosi di un limite in N e p, non dipende più esplicitamente da queste due grandezze.
Nel caso in cui si osservi il decadimento di un atomo, si è soliti procedere in questo modo: si suddivide il tempo di osservazione in intervalli (il che significa aver suddiviso gli infiniti tentativi in sottoinsiemi di infiniti tentativi) e si misura quante volte in ognuno di questi intervalli si verifica un successo. L'esperimento è ora praticamente suddiviso in più esperimenti minori da cui è possibile dedurre un numero medio frequentistico di successi. Per esempio:

Table 1: Decadimento di un atomo. Il tempo di osservazione è stato suddiviso in intervalli e per ogni intervallo è stato contato il numero di successi osservati.

# successi	0	1	2	3	4	5	6	7
# intervalli	1042	860	307	78	15	3	0	0
Poisson	1064	823	318	82	16	2	0.3	0.3

Il numero medio di eventi è:

$$\frac{1042 \cdot 0 + 860 \cdot 1 + 307 \cdot 2 + 78 \cdot 3 + 15 \cdot 4 + 3 \cdot 5 + 0 \cdot 6 + 0 \cdot 7}{1064 + 860 + 307 + 78 + 15 + 3 + 0 + 0} = 0.77$$

Da cui è possibile calcolare i valori sempre riportati nella tabella precedente.

3.6 Distribuzione uniforme

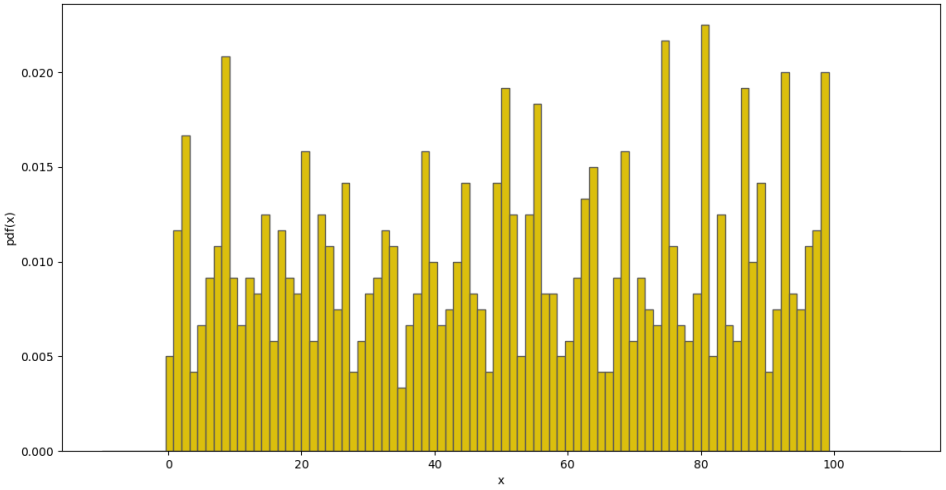


Figure 4: Distribuzione uniforme: $a = 0$, $b = 100$.

Una pdf di numeri che hanno tutti uguale probabilità di verificarsi è detta uniforme:

$$P(n, a, b) = \begin{cases} \frac{1}{b-a} & a \leqslant x \leqslant b \\ 0 & \text{altrove} \end{cases}$$

con:

$$E[n] = \frac{1}{2}(a + b)$$
$$V[n] = \frac{1}{12}(a + b)^2$$

Se una variabile è distribuita secondo una pdf $f(x)$, la sua cumulante è uniformemente distribuita. Intuitivamente è semplice perché basta vederla in questo modo: si immagini il grafico della pdf; ogni volta che si estrae un numero, questo cadrà in un punto casuale nell'area al di sotto della pdf, lasciando uno spazio casuale alla sua sinistra (che è il valore della cumulante)

3.7 Distribuzione Gaussiana e CLT

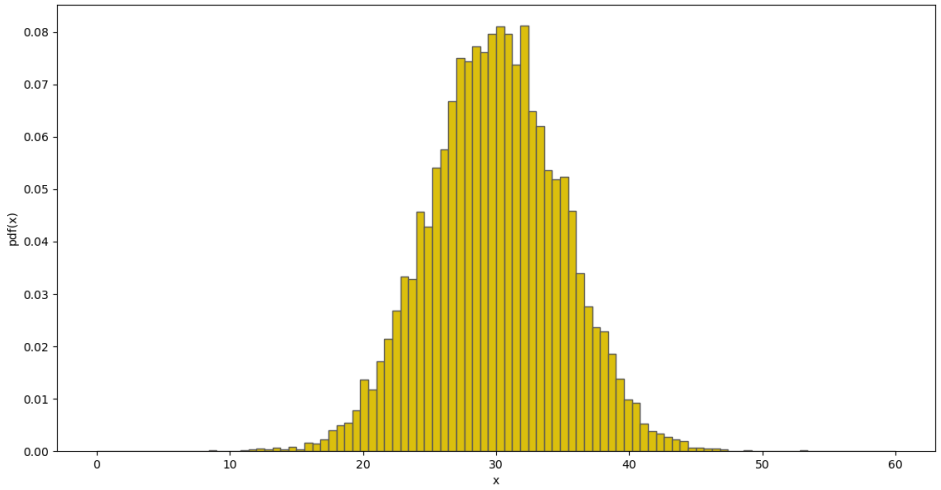


Figure 5: Distribuzione Gaussiana: $\mu = 30$, $\sigma = 5$.

La distribuzione Gaussiana (o normale) è definita come:

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

con:

$$E[x] = \mu$$
$$V[x] = \sigma^2$$

La error function, che è la cumulativa di questa pdf, è molto utile in laboratorio e i suoi valori sono tabulati.
Il teorema centrale del limite afferma che date n variabili casuali indipendenti distribuite con una pdf comune e varianze σ_i^2, nel limite in qui $n \to +\infty$, la somma di queste variabili segue un andamento gaussiano con valore medio la somma dei valori medi e varianza la somma delle varianze.
Ciò può essere sfruttato per generare numeri casuali distribuiti secondo una distribuzione normale.

Per grandi valori di μ (vale a dire qualche unità), la distribuzione di Poisson tende a quella Gaussiana con $\mu = \nu$ e $\sigma = \sqrt{\nu}$. Analogamente per $N \to +\infty$ la binomiale tende alla Gaussiana con $\mu = Np$ e $\sigma = \sqrt{Np(1-p)}$.

3.8 Distribuzione Gaussiana multivariata

Nel caso multidimensionale, la pdf per il vettore $\vec{x} = x_1 ... n_n$ è data da:

$$f(\vec{x}, \vec{\mu}, V) = \frac{1}{(2\pi)^{N/2} \mid V \mid^{1/2}} \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu})^t V^{-1} (\vec{x} - \vec{\mu})\right]$$

con $E[x_i] = \mu_i$ e $\text{cov}[x_i, x_j] = V_{ij}$

3.9 Media pesata

Quando si hanno misure con diversi errori, vanno combinate attraverso il concetto di media pesata:

$$E[x] = \frac{\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}$$

$$V[x] = \frac{1}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}$$

Ma non ha senso mediare valori che non sono compatibili!

3.10 Distribuzione di Breit-Wigner

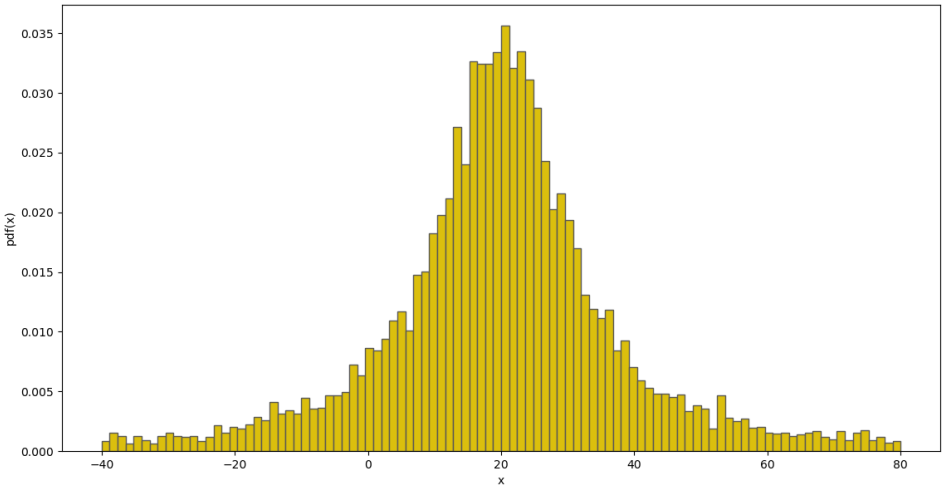


Figure 6: Distribuzione di Breit-Wigner: $x_0 = 20$, $\Gamma = 10$.

Esistono alcune distribuzioni che hanno momenti non ben definiti e che per questo si dicono "patologiche". Un esempio è la distribuzione di Breit-Wigner:

$$f(x, \Gamma, x_o) = \frac{1}{\pi} \cdot \frac{\Gamma/2}{\Gamma^2/4 + (x - x_0)^2}$$

Un caso particolare è quello in cui $x_0 = 0$ e $\Gamma = 2$, caso in cui è detta distribuzione di Cauchy:

$$f(x, 2, 0) = f(x) = \frac{1}{\pi} \cdot \frac{1}{1 + x^2}$$

Il valore medio e la varianza non sono definiti perché l'integrale è divergente. Conviene usare la moda e l'ampiezza a mezza altezza, che sono rispettivamente x_0 e Γ. Nella libreria *GSL*, la pdf è scritta in questo modo:

$$p(x) = \frac{1}{a\pi(1 + (x/a))^2} \qquad \Longrightarrow \qquad a = \Gamma/2$$

3.11 Distribuzione di Landau

Per una particella carica con $\beta = v/c$ che attraversa un materiale sottile di spessore d, la perdita di energia Δ segue la distribuzione di Landau:

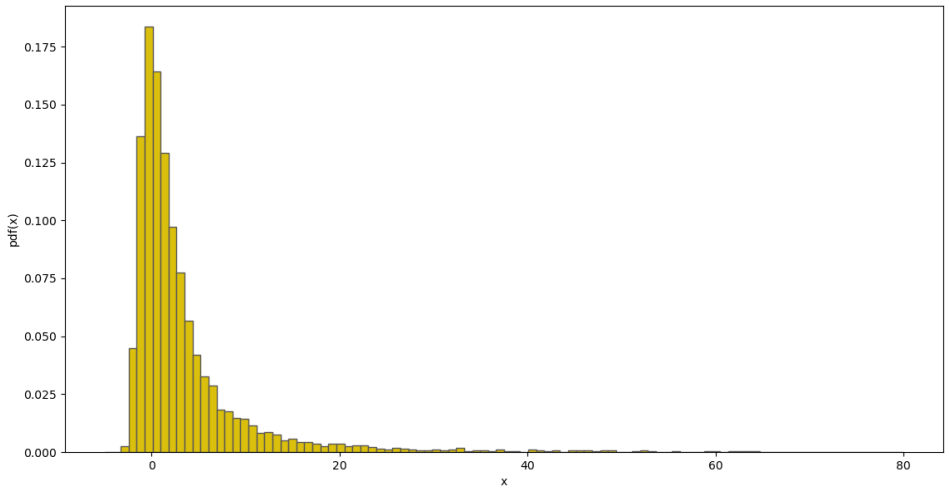


Figure 7: Distribuzione di Landau.

Ha una forma complicatissima che racchiude integrali, logaritmi… Anche in questo caso non si possono definire i momenti algebrici perché l'integrale diverge.

3.12 Distribuzione del chi-quadro

Date N grandezze distribuite ciascuna con una propria distribuzione Gaussiana, la somma dei loro quadrati segue la distribuzione χ^2. Formalmente è definita così:

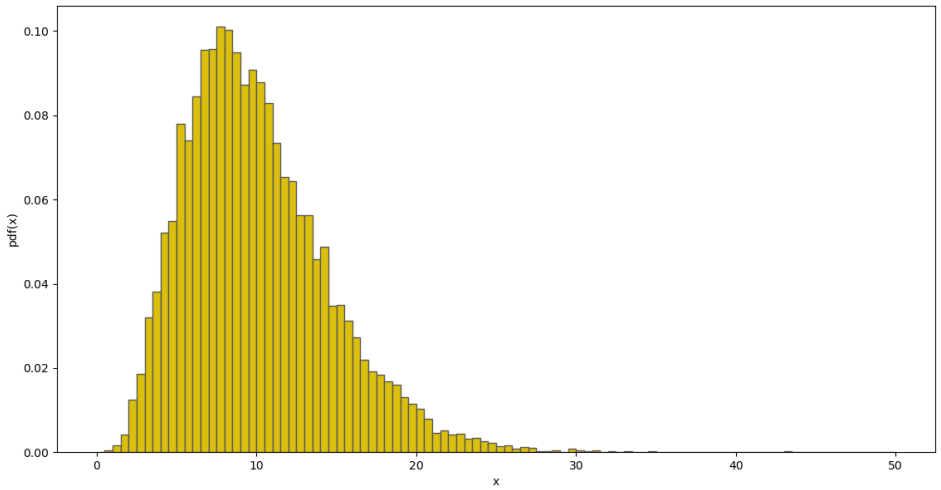


Figure 8: Distribuzione del χ^2: $n = 5$.

$$f(z,n)\frac{1}{2^{n/2}\Gamma(n/2)}z^{n/2-1}e^{-z/2}$$

$$E[z] = n$$
$$V[z] = 2n$$

dove z è la variabile e n è il numero di fradi di libertà.
Quando si fa un esperimento e si campiona $y(x)$ e poi si fittano i dati trovati con una funzione teorica $f(x)$, ciascun valore $y(x)$ si assume distribuito come una gaussiana attorno al suo valore vero, che assumiamo essere $f(x)$: dunque i residui, che sono la differenza $R(x) = y(x) - f(x)$, sono ancora una gaussiana, ma centrata in zero. Il chi quadro è definito come:

$$\sum_i \frac{[y(x_i) - f(x_i)]^2}{f(x_i)} \qquad \text{oppure} \qquad \sum_i \frac{[y(x_i) - f(x_i)]^2}{\sigma_i^2}$$

Ne consegue che il chi quadro segua appunto la distribuzione del chi quadro.

Nella libreria *GSL* la distribuzione χ^2 corrisponde alla distribuzione gamma con $a = n/2$ e $b = 2$.

3.13 Distribuzione esponenziale

$$f(x,\lambda) = \lambda e^{-\lambda x}$$

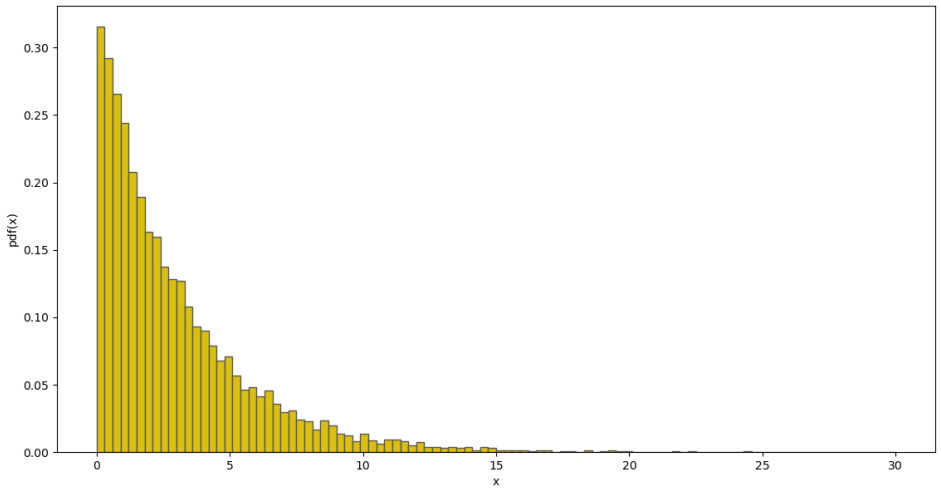


Figure 9: Distribuzione del esponenziale: $\lambda = 3$.

$$E[z] = \frac{1}{\lambda}$$
$$V[z] = \frac{1}{\lambda^2}$$

3.14 Distribuzione t di Student

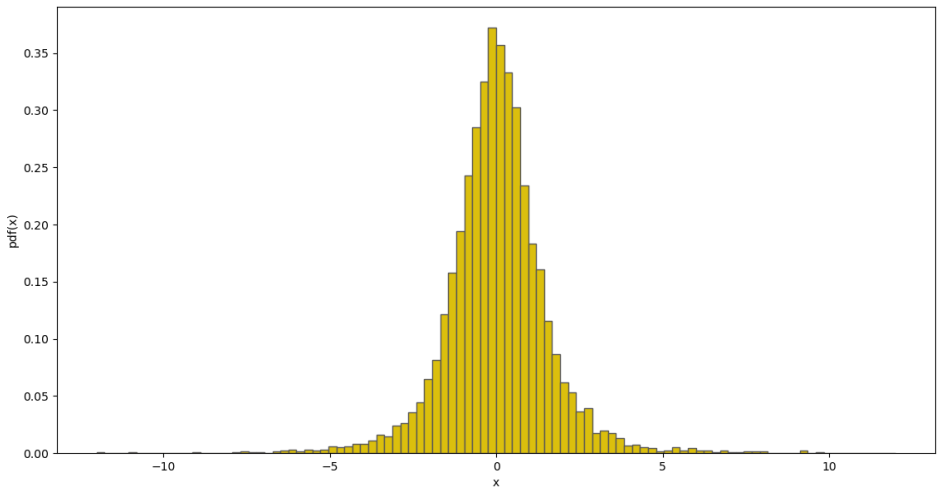


Figure 10: Distribuzione t di Student: $\nu = 3$.

È la distribuzione seguita dalla media di una popolazione gaussiana quando la si stima con un piccolo campione e senza conoscere la deviazione standard. Se y_1 è

distribuita come una Gaussiana e y_2 come un χ^2, se ν sono i gradi di libertà, allora x segue la t di Student:

$$x = \frac{y_1}{\sqrt{\frac{y_2}{\nu}}}$$

che è così definita:

$$f(x,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

$$E[z] = 0$$
$$V[z] = \begin{cases} \frac{\nu}{\nu-2} & \nu > 2 \\ \infty & \nu \leqslant 2 \end{cases}$$

3.15 Distribuzione di Fischer-Snedecor

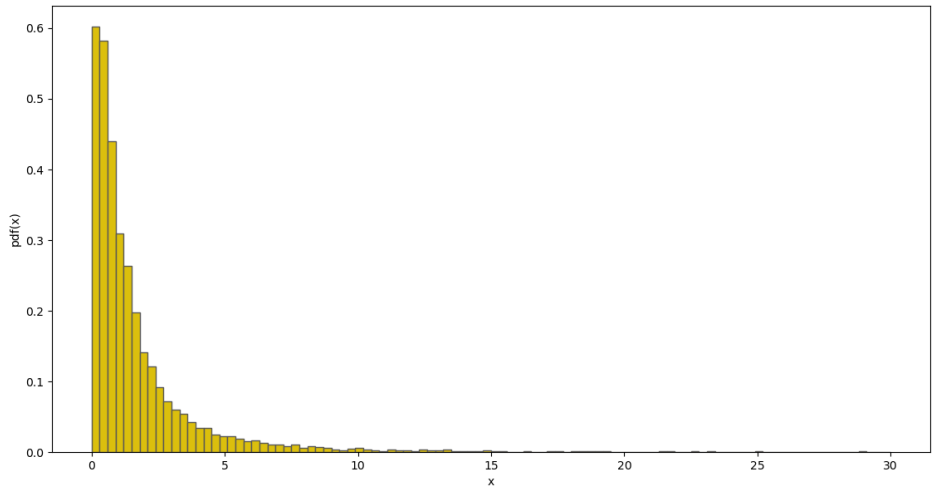


Figure 11: Distribuzione di Fischer: $n = 3$, $m = 4$.

Se si hanno due campioni $\vec{x}$ e $\vec{y}$ di variabili che seguono le rispettive Gaussiane, si può usare la distribuzione di Fisher-Snedecor per comparare le due varianze. Se nel primo caso le variabili sono n e nel secondo sono m, allora la distribuzione di Fisher con gradi di libertà $n-1$ e $m-1$ dà la distribuzione del rapporto:

$$\frac{S_x^2/S_y^2}{\sigma_x^2/\sigma_y^2} = \frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2}$$

con:

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_i)^2 \qquad , \qquad S_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \mu_i)^2$$

che quindi è il rapporto di due grandezze distribuite secondo il chi^2.
La definizione della pdf è complicata…

3.16 Funzione caratteristica

Si definisce funzione caratteristica di una variabile x distribuita secondo una $f(x)$, la trasformata di Fourier di quest'ultima:

$$\hat{f}(k) = E[e^{ikx}] = \int_{-\infty}^{+\infty} dx f(x) e^{ikx}$$

come per ogni trasformata, tutte le informazioni contenute nella funzione originaria sono contenute anche nella funzione caratteristica, perché per tornare alla prima è sufficiente calcolare la trasformata inversa:

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} dx \hat{f}(k) e^{-ikx}$$

la funzione caratteristica è utile per semplificare alcuni conti. Se $x_1 \dots x_N$ sono variabili casuali indipendenti:

$$\begin{aligned} z = \sum_{i=1}^{N} x_i \qquad \Longrightarrow \qquad \hat{f}_z(k) &= \int dx_1 \dots dx_N f_1(x_1) \dots f_N(x_N) e^{ik\sum_{i=1}^{N} x_i} = \\ &= \int dx_1 f_1(x_1) e^{ikx_1} \dots \int dx_N f_N(x_N) e^{ikx_N} = \\ &= \hat{f}_1(k) \dots \hat{f}_N(k) \end{aligned}$$

Inoltre vale anche che:

$$\frac{d^m}{dk^m}\hat{f}(k)\big|_{k=0} = \frac{d^m}{dk^m}\int_{-\infty}^{+\infty} dx f(x) e^{ikx}\big|_{k=0} = i^m \int_{-\infty}^{+\infty} dx f(x) e^{ikx} x^m\big|_{k=0} = i^m \mu_m = i^m E[x^m]$$

che è il momento algebrico di ordine m.
Per esempio, nel caso di due variabili indipendenti x e y gaussiane, si può notare subito che la loro somma è una gaussiana con $\mu = \mu_x + \mu_y$ e $\sigma^2 = \sigma_x^2 + \sigma_y^2$. Analogamente per la Poissoniana.
Inoltre è facile osservare quale sia il comportamento delle pdf nei vari limiti che abbiamo visto in precedenza: se si manda $N \to \infty$ mantenendo il valore medio costante nella funzione caratteristica di una binomiale, si ottiene la funzione caratteristica di una Poissoniana. Anche il teorema centrale del limite si può dimostrare in questo modo.

4 BPH

4.1 Statistica descrittiva

Si possono distinguere due tipi di analisi dei dati: "model independent" (statistica descrittiva) e "model dependent", che si basano su un modello teorico. In questo capitolo studiamo quelli del primo tipo.
Alcuni argomenti tipici di statistica descrittiva sono:

- test per stabilire se due datasets provengono dalla stessa distribuzione $f(x)$;
- test per stabilire la correlazione tra due datasets (test di ipotesi);
- metodi per determinare i momenti di una distribuzione;
- metodi per lo smoothing dei dati sperimentali.

4.1.1 Momenti di una distribuzione

Definire i momenti di una distribuzione ha senso quando gli eventi che la costituiscono hanno la tendenza ad agglomerarsi attorno ad un valore centrale. Se i dati sono discreti, si usano le seguenti definizioni:

Media campionaria: se n_i è la frequenza con cui si presenta ciascun valore x_j:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{j=1}^{N'} n_j x_j$$

Momento centrale di ordine r:

$$V_r = \frac{1}{N}\sum_{j=1}^{N'} n_j^r (x_j - \bar{x})^r$$

Esistono anche altri due valori "centrali", che nel caso continuo diventano: mediana:

$$\int_{-\infty}^{x_{\text{med}}} dx\, f(x) = \frac{1}{2}$$

moda: valore per cui $f(x)$ è massima, ovvero valore che si ripete con maggiore frequenza.
Se la pdf ha code molto estese, è possibile che gli integrali non convergano e questi valori non siano definiti. Per questo motivo la mediana è uno stimatore del valore centrale più robusto della media.

I momenti centrali definiscono il modo in cui i dati si distribuiscono attorno al valore centrale: quanto sono "diffusi". Il primo è la varianza (si noti la correzione di Bessel per cui $N \to N-1$ al denominatore):

$$V = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

La skewness (letteralmente "asimmetria") descrive quanto i valori siano distribuiti in modo disuniforme attorno al valore medio:

$$\gamma = \frac{1}{\sigma^3}E[(x - \bar{x})^3]$$

dove σ è la deviazione standard.

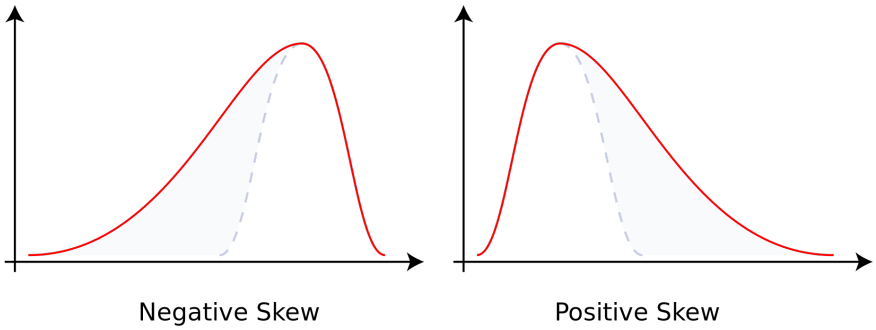


Figure 12: Skewness.

Quanto una pdf è più o meno piccata rispetto ad una gaussuana è dato dalla kurtosis ("curved", "arching"):

$$K = \frac{1}{\sigma^4}E[(x - \bar{x})^4] - 3$$

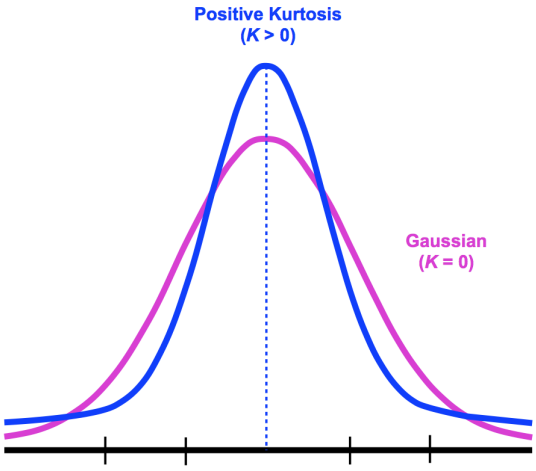


Figure 13: Kurtosis.

Esiste una stima per le deviazioni standard di questi parametri nel caso di distribizioni circa gaussiane:

$$V(\sigma^2) = \frac{2\sigma^4}{N}$$
$$V(\gamma) \approx \frac{15}{N}$$
$$V(K) \approx \frac{96}{N}$$

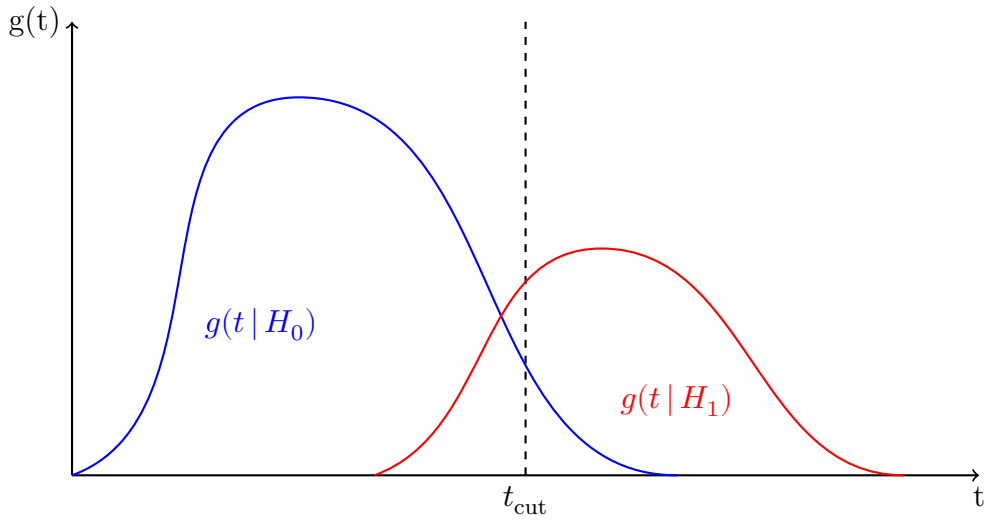
4.1.2 Smoothing dei dati

Lo smoothing dei dati si rende necessario quando i dati sono corrotti da un rumore casuale. Solitamente si attua una media su finestre che inglobano dati contigui. Fare una media, però, significa abbassare inevitabilmente il valore nei picchi, perché la maggior parte delle volte conservano l'area al di sotto del picco e la posizione, ma non l'altezza.
Uno dei più efficienti metodi di smoothing è il filtro di Savitsky-Golay.

Il segnale viene analizzato a gruppi di punti incentrati ciascuno in Y_i, con i che scorre su tutto l'array. Chiamiamo y_0 il punto centrale e Y_N e Y_{-N} gli estremi. Y_0 viene sostituito con un valore calcolato in un modo spiegato di seguito. Durante questo processo, i valori di Y_i non vengono sostituiti con f_i, bensì si crea un array parallelo che sarà poi quello definitivo smoothato.
I valori di Y_i si ottengono tramite un fit sui punti della finestra con un polinomio di grado arbitrario g: $P_g(j)$. Il polinomio viene poi valutato in zero e sostituito al valore di y_0.

4.1.3 Test di ipotesi

Supponiamo di voler dimostrare che una certa variabile casuale x segua una pdf $f(x)$: questa è detta ipotesi nulla H_0. Se $f(x)$ non dipende da alcun parametro, si parla di ipotesi semplice, altrimenti di dice composta. Oltre alla ipotesi nulla si possono avere una o più ipotesi alternative H_1, H_2…
Consideriamo il semplice caso in cui abbiamo una sola ipotesi alternativa H_1 che proponga a sua volta una pdf. Per valutare l'accordo tra i dati e un'ipotesi nulla si costruisce una statistica di test $t(x)$, che è una variabile che dipende da $\vec{x}$ che definisco per determinare se l'ipotesi nulla sia vera oppure no (vedi t_{cut} oppure la discrepanza…) e che segue a sua volta due pdf, una prevista da H_0 e una da H_1.



Si definisce 'significanza del criterio di test' α (mentre $(1-\alpha)$ è il 'livello di confidenza del criterio di test', o 'efficienza'):

$$\alpha = \int_{t_{\text{cut}}}^{+\infty} dt\, g(t \mid H_0)$$

mentre β è chiamato 'potenza del test' (mentre $(1-\beta)$ è detto 'purezza'):

$$\beta = \int_{-\infty}^{t_{\text{cut}}} dt\, g(t \mid H_1)$$

Si chiamano:

- errore di prima specie: rigezione di H_0 qualora questa sia vera (con relativa probabilità P_1);
- errore di seconda specie: accettazion di H_0 qualora questa sia falsa (con relativa probabilità P_2);

Per $t < t_{\text{cut}}$ deciso arbitrariamente, imponiamo che l'ipotesi nulla sia verificata. Ne consgue che $\alpha = P_1$ e $\beta = P_2$.
La scelta migliore di y_{cut} è quella che dà la massima purezza data una certa efficienza. Nel caso 1D lo si ottiene automaticamente (vedi esempio), altrimenti può essere complicato.

Facciamo un esempio in cui applichiamo il lemma di Neyman-Pearson.
Immaginiamo di avere i valori $\vec{x} = (x_1 ... x_N)$ che appartengono ad una distribuzione normale la cui varianza σ è nota e si deve distinguere tra due valori medi μ_0 e μ_1, cioé:

$$H_0 = [\mu = \mu_0] \qquad\qquad H_1 = [\mu = \mu_1]$$

A questo punto le pdf previste da H_0 e H_1 sono due gaussiane centrate ciascuna nel proprio valore medio. Secondo il lemma di cui sopra, dobbiamo calcolare la Likelihood, che è la produttoria su tutte le misure effettuate x_i della pdf prevista di un'ipotesi calcolata in x_i:

$$L(\vec{x}, \mu, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^N} \Pi_{i=1}^{N} N(x_i, \nu, \sigma)$$

dove con N si indica la distribuzione normale. Si tratta, cioè, della probabilità di avere ottenuto quelle misure secondo l'ipotesi considerata. Vorremo, quindi, che $L(H_0) >> L(H_1)$. A questo scopo si guarda r, parametro previsto dal lemma, che vale:

$$r = \frac{(L(\vec{x}) \,|\, H_0)}{(L(\vec{x}) \,|\, H_1)} \qquad \Longrightarrow \qquad \ln r = \ln L(\vec{x}, \mu_0, \sigma) - \ln L(\vec{x}, \mu_0, \sigma)$$

Che deve essere a sua volta molto grande. La regione in cui si deve accettare l'ipotesi nulla è infatti quella con $r > c$, dove c deve ancora essere valutato.

$$\ln r = R(\vec{x}) > \ln c \qquad \Longrightarrow \qquad \vec{x} > (\text{oppure} <)\, g(c) = t_{\text{cut}}$$

Per scegliere k, si impone che:

$$P_1 = \alpha = Pr(\vec{x} > (\text{oppure} <)\, t_{\text{cut}} \,|\, H_0)$$

Quindi ciò che può essere scelto arbitrariamente, alla fine dei conti, è α, che solitamente si impone $= 5\%$.

4.1.4 Discriminante lineare di Fisher

In che modo si possono definire $f(t \,|\, H_0)$ e $f(t \,|\, H_1)$? Si possono fare degli *ansatz* riguardo alla forma di t. Il modello di Fischer utilizza una funzione lineare:

$$t = \sum_{i=1}^{N} a_i x_i = \vec{a} \cdot \vec{x}$$

dove il vettore $\vec{a}$ è da determinare. Definiamo l'insieme dei valori medi e delle "varianze" delle variabili misurate come segue: $\mu_{k,i}$ è il valore medio della variabile i-esima secondo l'ipotesi k-esima:

$$\mu_{k,i} = \int_{-\infty}^{+\infty} dx_1 \dots dx_N \, x_i f(\vec{x} \,|\, H_k)$$

dove k può quindi essere 0 o 1; mentre:

$$(V_k)_{i,j} = \int_{-\infty}^{+\infty} dx_1 \dots dx_N \, (x_i - \mu_{k,i})(x_j - \mu_{k,j}) f(\vec{x} \,|\, H_k)$$

Si può dimostrare che, per funzioni gaussiane, la migliore statistica di test (ovvero che massimizza $1 - \beta$ per un dato a) è quella per cui:

$$\vec{a} = \frac{1}{w}(\vec{\nu}_0 - \vec{\nu}_1) \qquad \text{con} \qquad W_{i,j} = (V_0 + V_i)_{i,j}$$

In genere si introduce anche un offset:

$$t = a_0 + \sum_{i=1}^{N} a_i x_i$$

4.1.5 Reti neuronali

Si può dimostrare che se si usa il discriminante lineare di Fisher, allora dati i dati $\vec{x}$, la probabilità che sia giusta H_0 è:

$$P(H_0|\vec{x}) = frac11 + e^{-t}$$

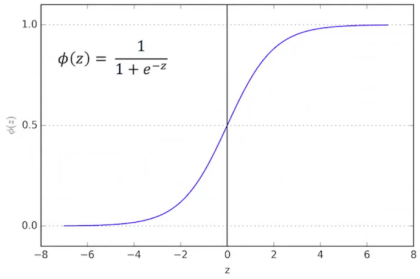


Figure 14: Logistic function.

che è la funzione logistica. Se le due pdf $f(\vec{x}|H_0)$ e $f(\vec{x}|H_1)$ non sono gaussiane, allora il discriminante lineare di Fisher non è più ottimale e si può generalizzare $t(\vec{x})$ con un caso speciale di Artificial Neural Network (ANN).
Supponiamo di prendere

$$t(\vec{x}) = s_0 \left(a_0 \sum_{i=1}^{N} a_i x_i \right)$$

con s detta funzione di attivazione e a_0 detta soglia. Siccome la sigmoide è monotona, questa ANN è equivalente ad un test lineare.