

Statistica

Progetto Telerilevamento - Liceo Manzoni a.s.
2020/2021

Giuliana Arrigo

Che cos'è la Statistica?

La Statistica è quel settore della matematica che si occupa dello studio di fenomeni collettivi allo scopo di stimare una o più caratteristiche della popolazione che si studia, di verificare ipotesi sulle stesse, di esplorare ed eventualmente verificare quali relazioni esistono tra fenomeni d'interesse.

Ricordiamo la terminologia specifica:

Popolazione: Insieme degli elementi che viene preso in considerazione per studiarne uno o più caratteristiche dette **caratteri**

Campione: Sottoinsieme della popolazione opportunamente selezionato

Unità statistica: ogni singolo elemento della popolazione o del campione

Dato statistico: il risultato del rilevamento della modalità assunta da un carattere su una unità statistica

Le fasi di una indagine statistica

1. Definizione degli obiettivi dell'indagine
2. Definizione del disegno dell'indagine
3. Acquisizione dei dati
4. Registrazione
5. Revisione e validazione
6. Presentazione e utilizzazione dei risultati
7. Diffusione

Come si suddivide

La Statistica si distingue in :

Statistica descrittiva: se l'indagine è svolta sull'intera popolazione

Accurata, ma molto dispendiosa in termini di costi e di tempi. Inoltre non sempre è possibile attuarla.

Statistica induttiva: se l'indagine è svolta su un campione rappresentativo

Se le unità sono selezionate in modo casuale, i risultati possono essere estesi a tutta la popolazione grazie a tecniche probabilistiche ed è possibile stimare la bontà dei risultati.

La probabilità

La probabilità si occupa di creare dei modelli matematici che permettano di descrivere un fenomeno aleatorio.

Definiamo:

Fenomeno aleatorio: un fenomeno che anche se ripetuto più volte e nelle stesse condizioni può dar luogo a esiti diversi e non prevedibili

Spazio campionario: l'insieme di tutti i possibili esiti di un fenomeno aleatorio

Evento aleatorio: ogni sottoinsieme dello spazio campionario

Variabile aleatoria: Una grandezza che può assumere un valore non determinato a priori a seguito del verificarsi di un evento aleatorio

Definizione di probabilità

Definizione classica (a priori):

$$p(E) = \frac{\textit{numero dei casi favorevoli}}{\textit{numero dei casi possibili}}$$

Definizione frequentistica (a posteriori):

$$p(E) = \frac{\textit{numero delle prove favorevoli}}{\textit{numero delle prove effettuate}}$$

La distribuzione di probabilità di una variabile aleatoria

Abbiamo definito in precedenza **variabile aleatoria**, X , una grandezza che può assumere un valore non determinato a priori a seguito del verificarsi di un evento aleatorio. Pertanto si può considerare come una funzione definita sullo spazio campionario.

Variabile aleatoria **discreta** se lo spazio campionario è finito o numerabile

Variabile aleatoria **continua** se lo spazio campionario è un intervallo reale limitato o illimitato

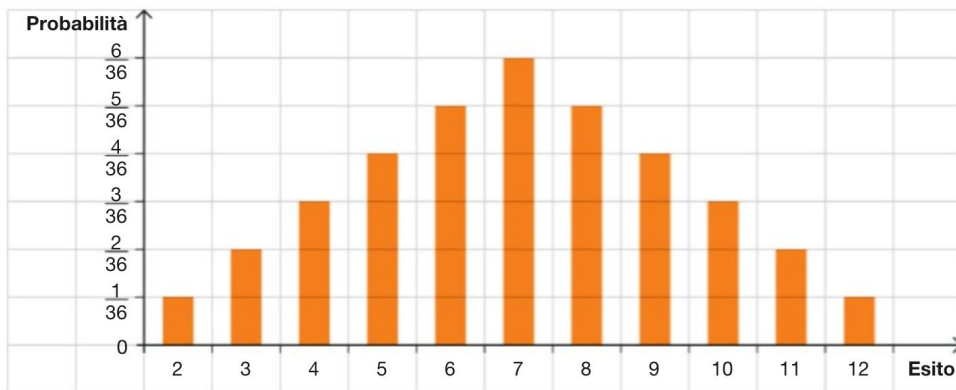
Si definisce **distribuzione di probabilità di una variabile aleatoria discreta** una relazione che stabilisce una corrispondenza tra i valori possibili della variabile

$\{x_1, x_2, \dots, x_n\}$ e la loro probabilità di accadere $P_i = P\{X = x_i\}$

Esempio distribuzione di probabilità discreta

Distribuzione della somma dei numeri presenti sulle facce di due dadi lanciati contemporaneamente

Esito	2	3	4	5	6	7	8	9	10	11	12
P	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



La media, varianza e deviazione standard

La **media** di una variabile aleatoria X è data da:

$$M(X) = x_1 \cdot P_1 + x_2 \cdot P_2 + \dots + x_n \cdot P_n$$

La media di una variabile aleatoria si chiama anche **speranza matematica**

La **varianza** di X è data da:

$$\text{Var}(X) = \sigma_x^2 = (x_1 - m)^2 \cdot P_1 + (x_2 - m)^2 \cdot P_2 + \dots + (x_n - m)^2 \cdot P_n$$

La **deviazione standard**, o scarto quadratico medio, da:

$$\sigma(X) = \sqrt{(\sigma_x^2)} = \sqrt{(x_1 - m)^2 \cdot P_1 + (x_2 - m)^2 \cdot P_2 + \dots + (x_n - m)^2 \cdot P_n}$$

Le distribuzioni di probabilità discrete più comuni

La distribuzione uniforme discreta: $P(X = x_i) = \frac{1}{n}$

La distribuzione di Bernoulli: $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$

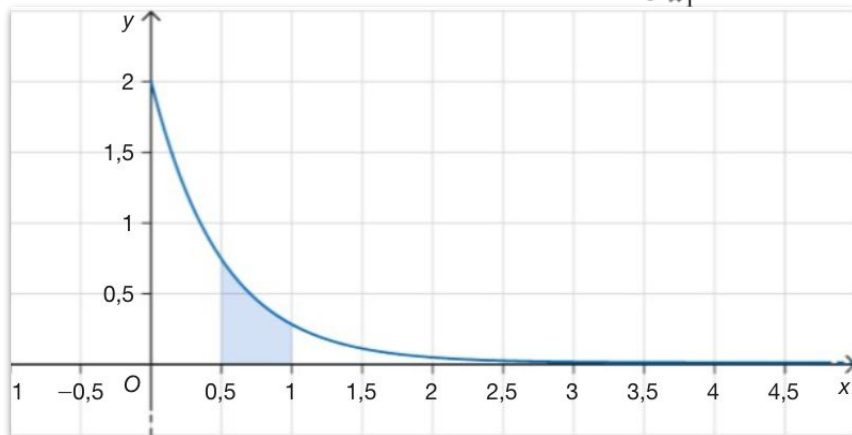
La distribuzione geometrica: $P(X = x) = (1 - p)^{x-1} \cdot p$

La distribuzione di Poisson: $P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$

Le distribuzioni di probabilità continue

Nel caso di una variabile aleatoria continua non ha senso considerare la probabilità che si verifichi un evento elementare, ma piuttosto la probabilità che la variabile aleatoria assuma un valore compreso tra due valori dati

$$P(x_1 \leq X \leq x_2) = P(x \in [x_1; x_2]) = \int_{x_1}^{x_2} f(x) dx$$

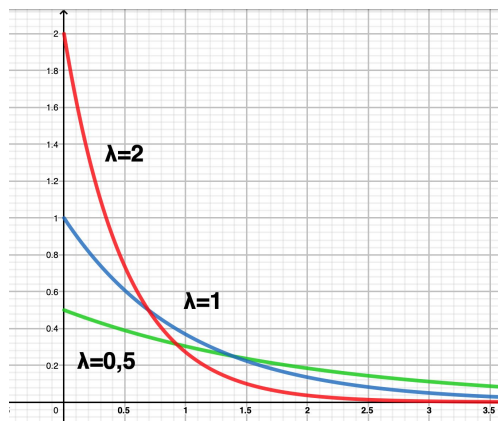


La distribuzione esponenziale

E' una distribuzione statistica che modella il tempo che intercorre tra gli eventi. La sua funzione è:

$$y = \lambda e^{-\lambda x}$$

dipenda da un parametro, lambda, che è proporzionale alla rapidità con cui gli eventi si ripresentano



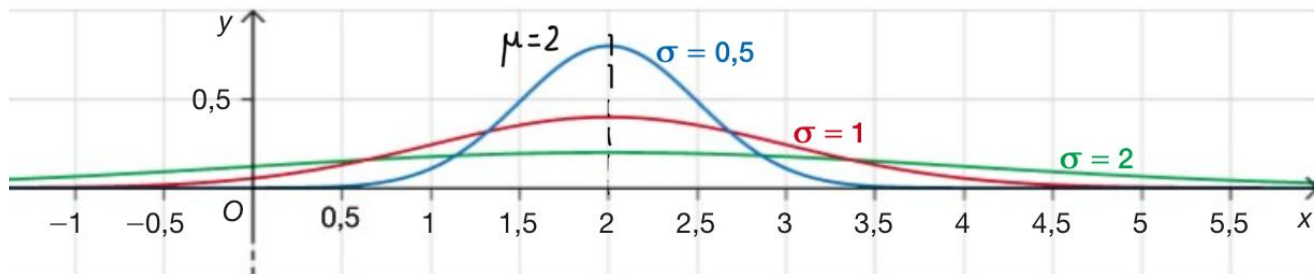
La distribuzione gaussiana

E' la più nota tra le distribuzioni. Si tratta di una curva a simmetria centrale dalla caratteristica forma a campana.

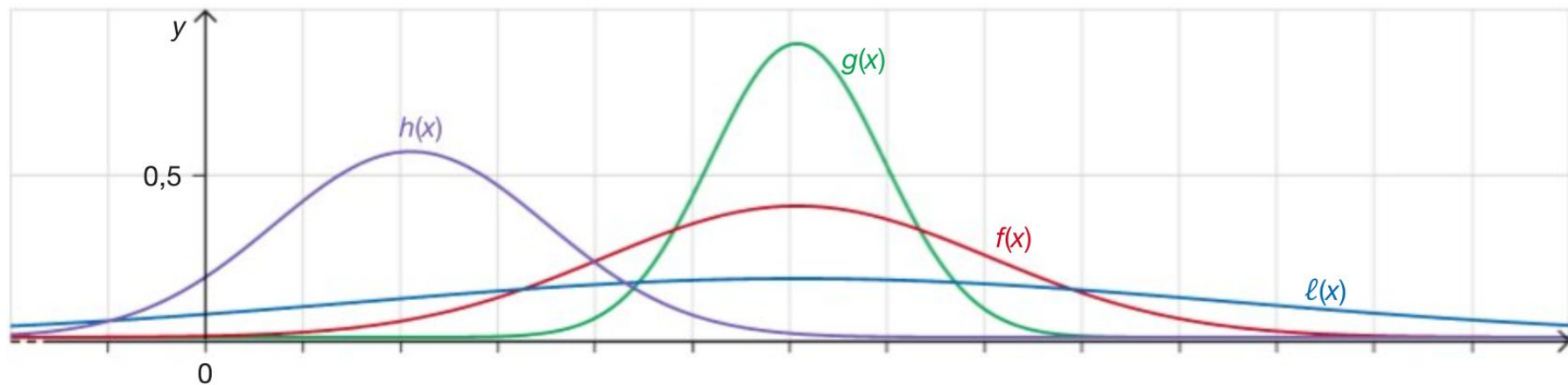
La sua equazione è: $y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma}}$ dipende da due parametri :

μ è la media della distribuzione normale

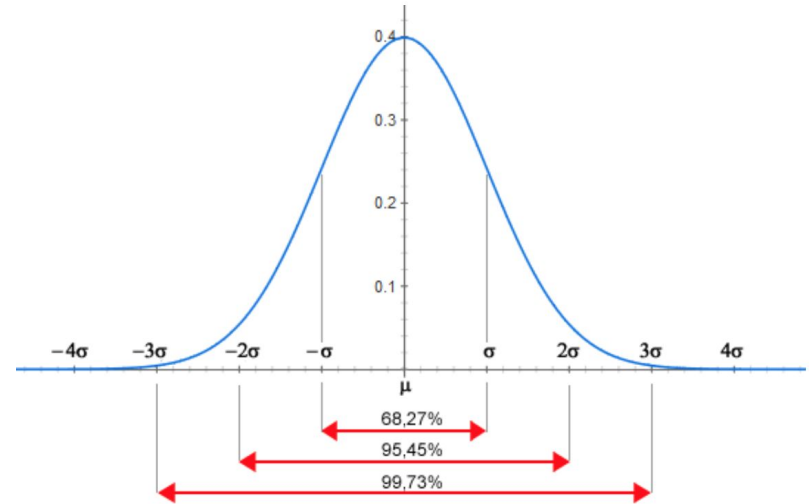
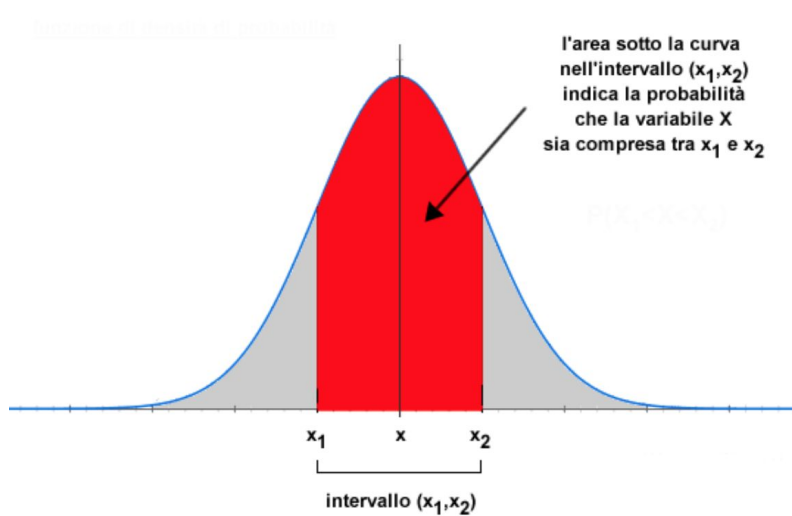
σ è la deviazione standard della distribuzione



Prova tu ad ordinare i valori di μ e σ



Altra caratteristica della distribuzione gaussiana



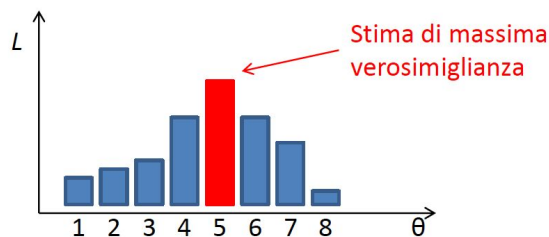
La massima verosimiglianza

La verosimiglianza misura la bontà con cui un insieme di dati “sostiene” una particolare distribuzione e si definisce come:

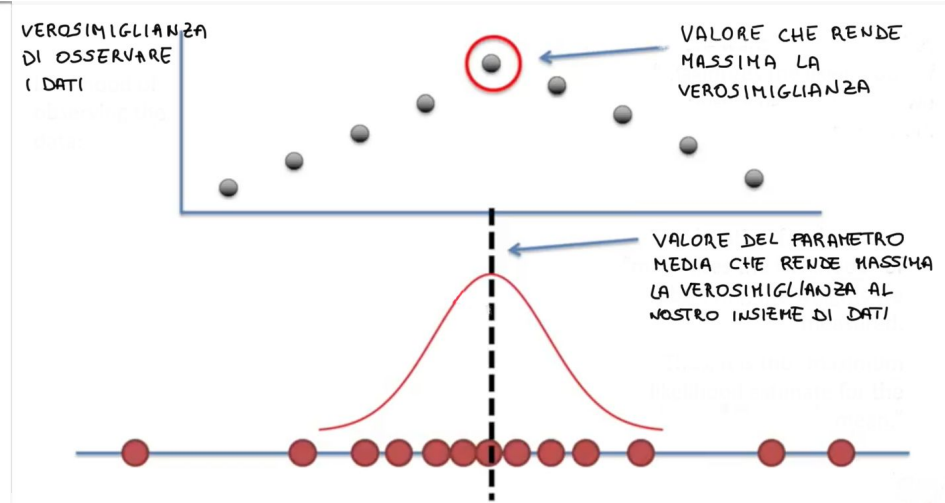
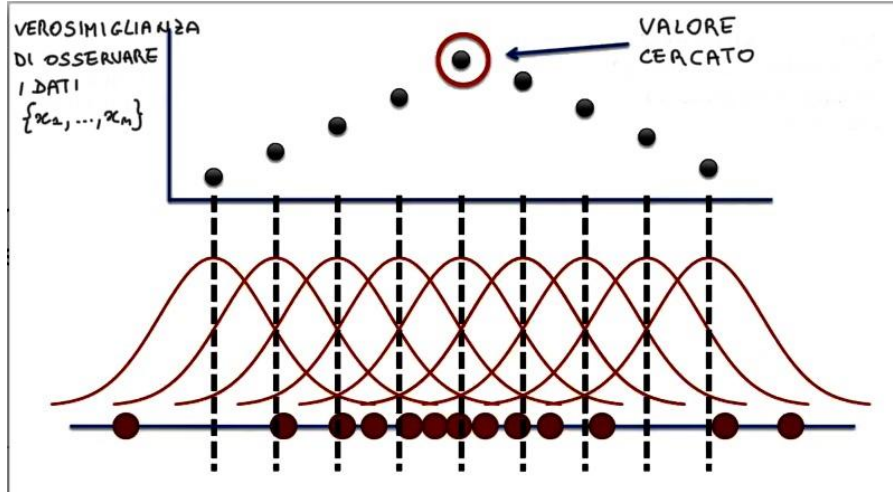
$$L(\text{distribuzione} \mid \text{dati}) = pr(\text{dati} \mid \text{distribuzione})$$

La distribuzione è identificata dal o dai suoi parametri caratteristici.

La verosimiglianza da sola non fornisce informazioni se non quando è confrontata con i valori di verosimiglianza per altri possibili parametri. La massima verosimiglianza ci aiuterà a trovare la migliore distribuzione che rappresenta un insieme di dati.

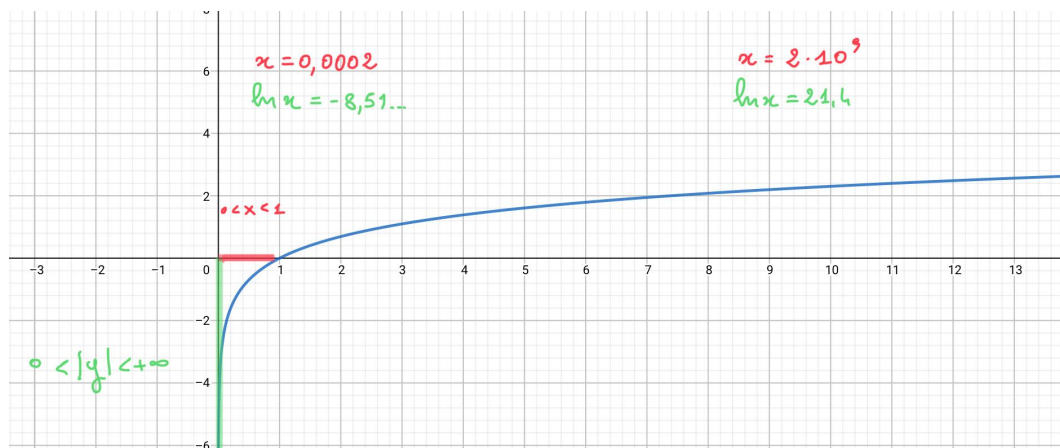


Massima verosimiglianza e distribuzione gaussiana

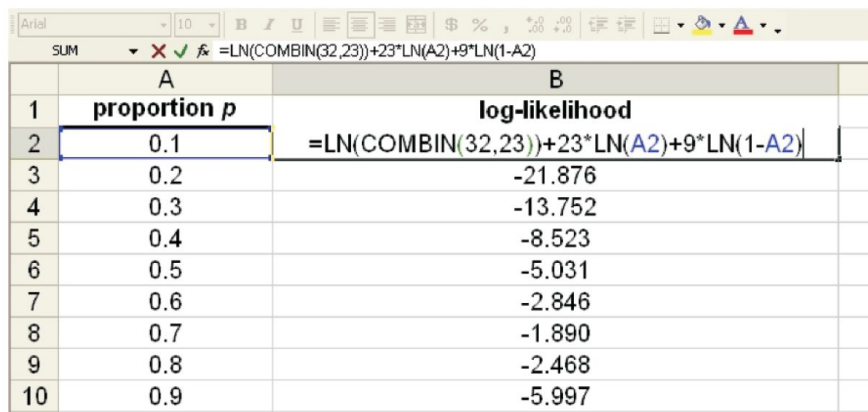


Log-verosimiglianza

Grazie alle caratteristiche della funzione logaritmo, spesso alla funzione di verosimiglianza si sostituisce il suo logaritmo: $\ln L(\mu, \sigma | x_1, \dots, x_n)$



Come si calcola il massimo della funzione L?



	A	B
1	proportion p	log-likelihood
2	0.1	=LN(COMBIN(32,23))+23*LN(A2)+9*LN(1-A2)
3	0.2	-21.876
4	0.3	-13.752
5	0.4	-8.523
6	0.5	-5.031
7	0.6	-2.846
8	0.7	-1.890
9	0.8	-2.468
10	0.9	-5.997

