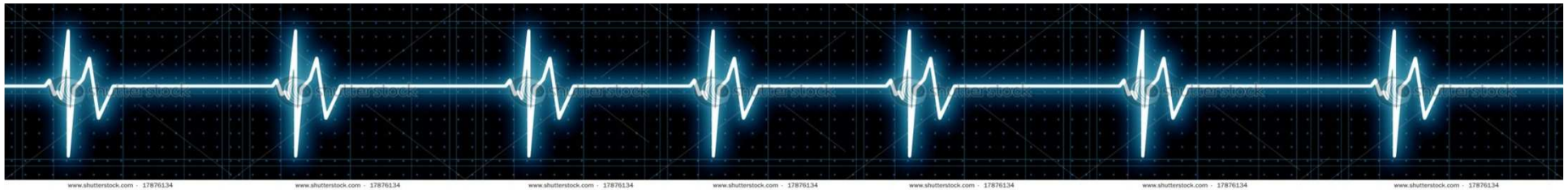
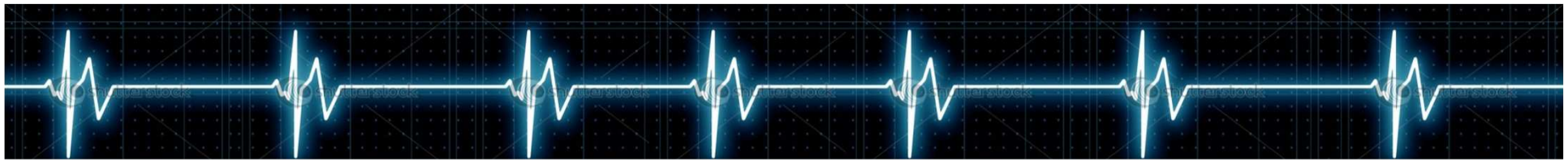


Variabili aleatorie



Esistono grandezze deterministiche?

la misura quasi sempre comporta un certo grado di approssimazione….

Esistono grandezze aleatorie?

l'aleatorietà è spesso legata al nostro grado di ignoranza del fenomeno……

Useremo allora:
l'approccio deterministico per studiare

- fenomeni "semplici"
- singole osservazioni

l'approccio aleatorio per studiare

- fenomeni "complessi"
- fenomeni che coinvolgono un numero elevato di realizzazioni

Variabile aleatoria

È il modello adatto a descrivere un esperimento "governato dal caso", quindi con qualche elemento di causalità, in cui quello che si osserva è un esito numerico, che non è quindi completamente prevedibile a priori. Ciò non vuol dire che l'esito sia completamente imprevedibile, ad es. nel lancio del dado so quali sono i possibili esiti, e so anche che, se il dado non è truccato, questi esiti sono equiprobabili.

- Statistica descrittiva
- Statistica inferenziale
- Calcolo delle probabilità

forniscono degli strumenti per la caratterizzazione e la valutazione quantitativa di grandezze aleatorie

Alcuni elementi di statistica descrittiva

Statistica (= studio delle cose dello Stato) si occupa dello studio di popolazioni, cioè di aggregati di individui, dove per individuo non si intende necessariamente un essere vivente o un individuo materiale, es. una popolazione può essere un insieme di misure.

La statistica descrittiva è la branca della Statistica che studia i criteri di rilevazione, di classificazione e di sintesi delle informazioni relative a una popolazione oggetto di studio. Tra i suoi obbiettivi:

- Raccogliere le informazioni sulla popolazione o su una parte di essa (campione), e sintetizzarli attraverso strumenti grafici (diagrammi a barre, istogrammi, boxplot) e indici (media, varianza, percentili, ecc.)
- Eseguire indagini di tipo comparativo
- Verificare l'adattamento dei dati empirici ad un modello teorico

MEDIA

ESEMPIO *CR= Carbo ratio (rapporto tra grammi di carboidrati ingeriti e le unità di insulina da iniettare in un soggetto diabetico)*

subj	CR
	(g/U)
1	13,4
2	14,0
3	14,8
4	52,0
5	20,0
6	16,6
7	23,3
8	17,4
9	25,2
10	25,0
11	30,0
12	21,8
13	18,3
14	21,0
15	18,0
16	18,0
17	28,2
18	16,6
19	27,0
20	37,0

Per dare una descrizione "media" di posizione

MATLAB
`m=mean(x)`

Calcola la media dei valori nel vettore x

$$\text{Media campionaria} = x_m = \frac{1}{N}\sum_{i=1}^{n} x_i = 22.9$$

$$\text{Mediana} = 20.5$$

Essendo N pari, la mediana è stata calcolata mediando i due elementi che si trovano in posizione centrale, dopo che i dati sono stati ordinati in senso crescente (vd più avanti la slide sui percentili)

Se N fosse dispari, si sceglierebbe il valore che si trova in posizione centrale

Media o mediana?

subj	CR
	(g/U)
1	13,4
2	140
3	14,8
4	52,0
5	20,0
6	16,6
7	23,3
8	17,4
9	25,2
10	25,0
11	30,0
12	21,8
13	18,3
14	21,0
15	18,0
16	18,0
17	28,2
18	16,6
19	27,0
20	37,0

La mediana è meno sensibile agli outliers

$$\text{Media campionaria} = x_m = \frac{1}{N}\sum_{i=1}^{n} x_i = 29.2$$

cambia

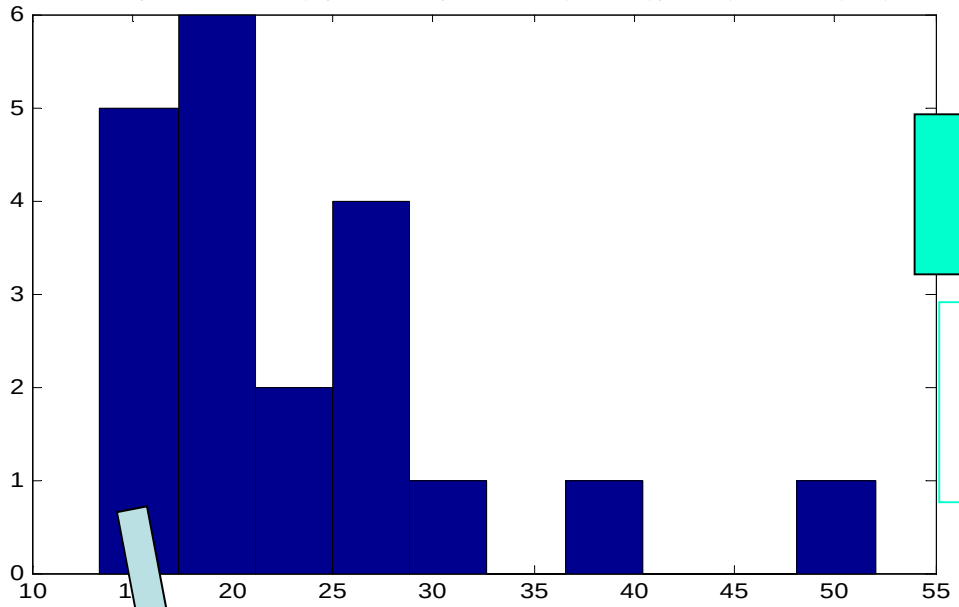
$$\text{Mediana} = 20.5$$

NON cambia

ISTOGRAMMA

subj	CR
	(g/U)
1	13,4
2	14,0
3	14,8
4	52,0
5	20,0
6	16,6
7	23,3
8	17,4
9	25,2
10	25,0
11	30,0
12	21,8
13	18,3
14	21,0
15	18,0
16	18,0
17	28,2
18	16,6
19	27,0
20	37,0

Per descrivere la variabilità



MATLAB
hist(x)

Valuta l'istogramma dei dati contenuti nel vettore x

Il range (da 13.4 a 52) viene diviso in 10 intervalli.

Il primo intervallo va da 13.4 a 17.26

5 elementi hanno un valore compreso tra gli estremi di questo intervallo

VARIANZA

subj	CR
	(g/U)
1	13,4
2	14,0
3	14,8
4	52,0
5	20,0
6	16,6
7	23,3
8	17,4
9	25,2
10	25,0
11	30,0
12	21,8
13	18,3
14	21,0
15	18,0
16	18,0
17	28,2
18	16,6
19	27,0
20	37,0

Per quantificare la variabilità

$$\text{Varianza campionaria} = s_x^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - x_m)^2 = 83.08$$

$$\text{Dev standard} = \sqrt{\text{varianza}} = 9.115$$

$$\text{Coeff di variazione} = \frac{\text{dev standard}}{\text{media}} \approx 11\%$$

MATLAB
`x=std(x)`

Calcola la SD dei valori nel vettore x

Media e dev standard sintetizzano:

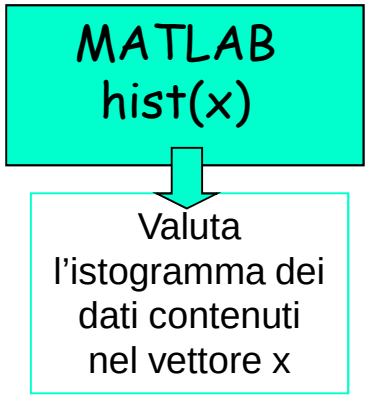
il valore attorno al quale si dispongono i vari elementi del campione

quanto i vari elementi sono dispersi attorno al valore medio

A cosa serve l'istogramma?

La lettura dell'istogramma consente di rispondere alle seguenti domande

- Quali sono gli intervalli con la maggiore (minore) densità?
- L'istogramma è unimodale?
- L'istogramma è simmetrico?
- La densità ha un andamento regolare (monotono crescente/decrescente, prima crescente poi decrescente, ecc.)?
- Ci sono classi sparse?



Riassumendo

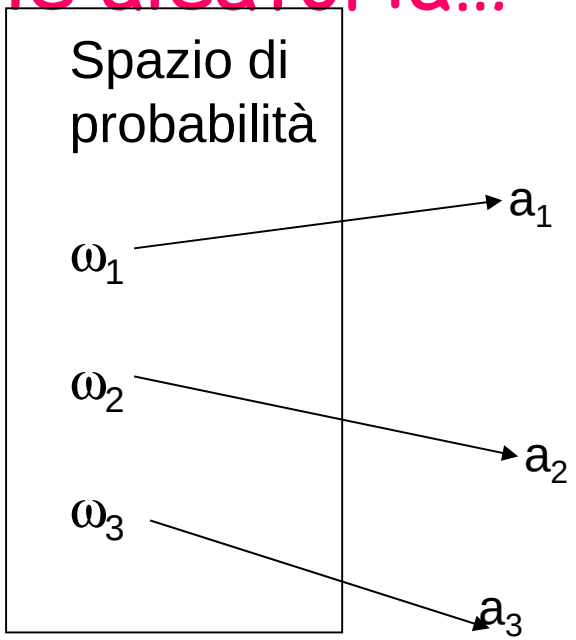
La statistica descrittiva si occupa dei dati osservati, prescindendo sia da qualsiasi modello probabilistico che descriva il fenomeno, sia dal fatto che l'insieme dei dati provenga da un campione estratto da una popolazione più vasta, o coincida con la popolazione

Abbiamo infatti visto come attraverso indici di posizione, dispersione e correlazione sia possibile descrivere in modo sintetico:

- un insieme (anche molto numeroso) di osservazioni numeriche
- la relazione tra due insiemi di osservazioni numeriche

Tornando al concetto di variabile aleatoria…

Spesso interessa descrivere un esperimento "governato dal caso" con un modello probabilistico, cioè caratterizzando la probabilità che l'esito assuma certi valori. Definiamo come v.a. una funzione che associa ad ogni campione dello spazio di probabilità (evento elementare) un valore numerico, appartenente ad un insieme che può essere discreto o continuo.



v.a. discreta

Può assumere un numero finito di valori

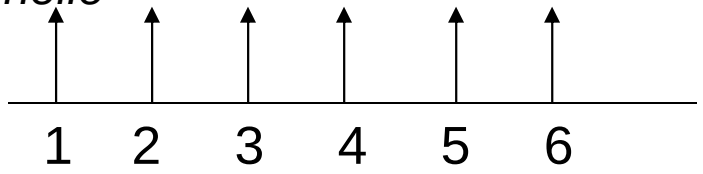
A = [a_1 a_2 a_3 …….. a_n] ciascuno associato ad una probabilità

$P[a_1]$ $P[a_2]$ $P[a_3]$ $P[a_n]$ con $\sum_{i=1}^{n} P[a_i] = 1$

Esempio: lancio del dado. Il numero che appare sulla faccia sup del dado è una v.a. che assume valori nello spazio discreto A

A = [1 2 3 4 5 6]

P[1] = P[2] = P[3] = P[4] = P[5] = P[6] = 1/6



Segue dalla definizione classica di probabilità, che si applica ad esperimenti elementari ritenibili equiprobabili: "la probabilità di un evento è il rapporto tra il numero di casi favorevoli e il numero totale di casi possibili, purchè quest'ultimi siano ugualmente possibili".

Dalla conoscenza di P si può calcolare la probabilità di qualsiasi evento

$$X \in A : P(X) = \sum_{a_i \in X} P(a_i)$$

Esempio: esito del lancio del dado

Prob(esito pari)= prob(2,4,6)=3/6

Esercizio: sia X la variabile che indica la somma dei punteggi di due dadi. Caratterizzare la v.a. X con la sua ddp

v.a. continua

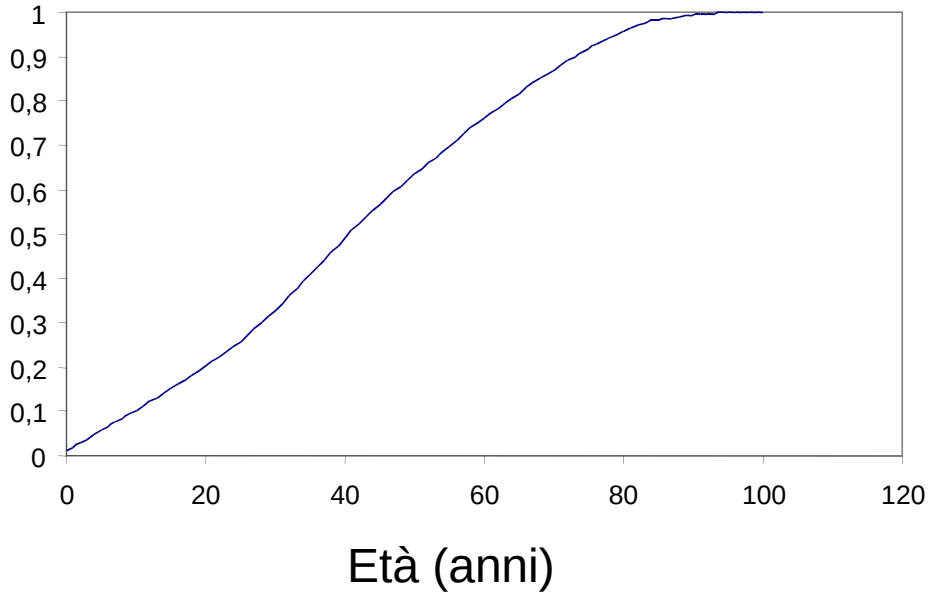
Assume valori appartenenti ad un insieme continuo. La definizione di probabilità nel continuo è più delicata che nel discreto. Conviene partire con:

Funzione di distribuzione di v.a. continua

$$F_A(a) = P[A \leq a]$$

Proprieta': $0 \leq F_A(a) \leq 1$

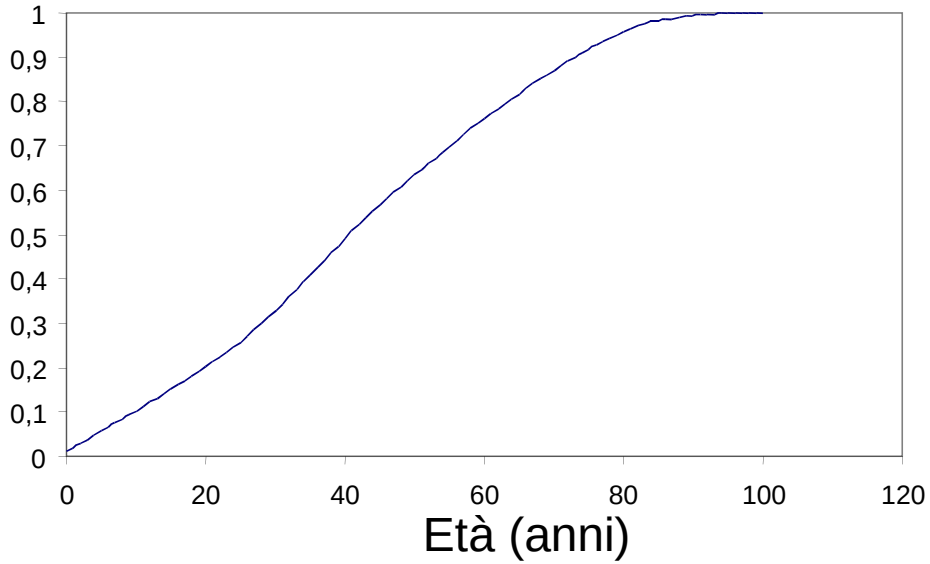
$F_A(a+h) \geq F_A(a)$ se $h>0$



Esempio: Si sceglie a caso un individuo di nazionalità italiana e si associa ad esso la sua età. Questa è una v.a. continua

La funzione distribuzione ha una importanza fondamentale per la descrizione statistica della v.a. A. Infatti dalla conoscenza della distribuzione si può ricavare la probabilità che A appartenga ad un generico insieme dell'asse reale

$$P[a_1 \leq A \leq a_2] = P[A \leq a_2] - P[A \leq a_1] = F_A(a_2) - F_A(a_1)$$



Esempio: *età di un individuo di nazionalità italiana*

(dati ISTAT http://demo.istat.it/pop2005/index.html)

$$P[10 \leq \text{età} \leq 20] = P[\text{età} \leq 20] - P[\text{età} \leq 10] = F_A(20) - F_A(10) = 0.099$$

$$P[20 \leq \text{età} \leq 30] = P[\text{età} \leq 30] - P[\text{età} \leq 20] = F_A(30) - F_A(20) = 0.1275$$

$$P[80 \leq \text{età} \leq 90] = P[\text{età} \leq 90] - P[\text{età} \leq 80] = F_A(90) - F_A(80) = 0.036$$

Densità di probabilità (ddp)

Si definisce densità di probabilità la derivata della funzione di distribuzione

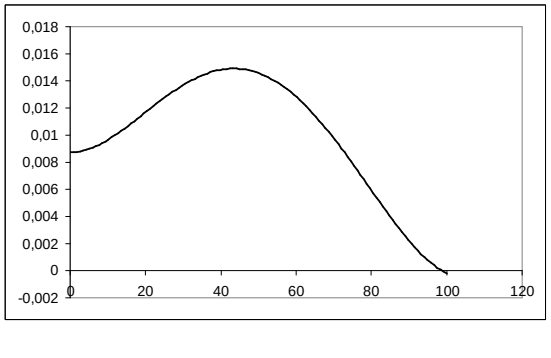
$$f_A(a)=d\,F_A(a)/da$$

Proprieta':

$f_A(a) \geq 0$

$$\int_{-\infty}^{a_1} f_A(a)da = F_A(a_1)$$

$$\int_{-\infty}^{+\infty} f_A(a)da = 1$$



Anche dalla conoscenza di $f_A(a)$ si può calcolare la probabilità che A appartenga ad un certo insieme dell'asse reale

$$P[a_1 \leq A \leq a_2] = \int_{a_1}^{a_2} f_A(a)da$$

Differenza tra probabilità e ddp

Mentre la P di una v.a. discreta in un certo valore a_1 rappresenta la probabilità che essa assume il valore a_1

la ddp di una v.a. continua in un certo valore a_1 NON rappresenta la probabilità che essa assume il valore a_1

Infatti la ddp di una v.a. continua NON è una probabilità ma una densità di probabilità, ovvero solo il suo integrale su un intervallo ha il significato di probabilità

Intervalli di confidenza

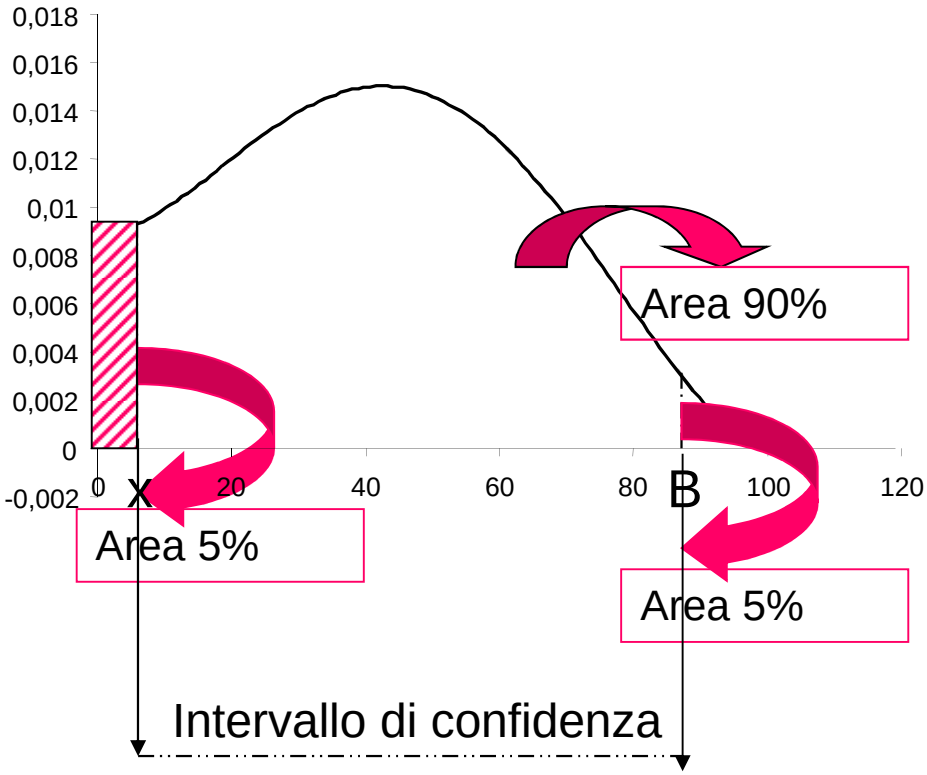
Intervalli di valori in cui, con una certa probabilità, cadono i valori della v.a:
Si valutano a partire dalla ddp (o dalla distribuzione):

Si fissa un livello di significatività (ad es. α=90%)

Si valutano gli estremi x e y all'interno dei quali cade una percentuale α di valori

$$\int_x^y f_A(a)da =$$

$$= F_A(y) - F_A(x) = \alpha$$



Esempio: età di un individuo di nazionalità italiana

Interv di conf	*95%*	*2-83 anni*
	90%	*5-79 anni*
	80%	*10-73 anni*

Media o aspettazione di una v.a. discreta

$$E[A] = \sum_{a_i \in A}^{n} a_i P[a_i]$$

Peso i valori a_i con la loro probabilità

Media o aspettazione di una v.a. continua

$$E[A] = \int_{-\infty}^{+\infty} a \cdot f_A(a) da$$

Esempio: *esito del lancio del dado*

Media = 1/6(1+2+3+4+5+6)=3.5* (NON è un valore assunto dalla v.a.)

Esempio: età della popolazione italiana

Media = 42

Esercizio: sia X la variabile che indica la somma dei punteggi di due dadi. Verificare che la sua media è 7

Proprietà

L'aspettazione è un operatore lineare:

Se $A_1 \ldots A_n$ sono v.a. discrete o continue, per ogni b_0, $b_1 \ldots . b_n$:

$E[b_0+b_1A_1+\ldots\ldots b_nA_n]=b_0+b_1E[A_1]+\ldots . b_nE[A_n]$

Teorema dell'aspettazione

Valore atteso (o media) di una funzione di v.a.

Discreta

$$E[g(A)] = \sum_i g(a_i)P[a_i]$$

Continua

$$E[g(A)] = \int_R g(a)f_A(a)da$$

L'utilità di queste formule è evidente: per calcolare la media di una funzione g(A) non serve conoscere la ddp di g(A) ma solo quella di A

Momento di ordine k di una v.a.

$$m_A(k) = E[A^k]$$

Momento di ordine 1: aspettazione, o valor medio

$$m_A = E[A]$$

Momento di ordine 2: potenza statistica

$$M_A = E[A^2]$$

Varianza (misura della dispersione attorno al valor medio)

$$\sigma_A^2 = E[(A - m_A)^2]$$

Deviazione Standard (SD) è la radice quadrata della varianza, pari a σ_A

Si noti l'analogia tra varianza della v.a. e varianza campionaria

Esercizio: esito del lancio del dado. Verificare che SD=1.70

Esercizio: sia X la variabile che indica la somma dei punteggi di due dadi. Verificare che la sua media è 7 senza usare la ddp di X ma direttamente dal teorema dell'aspettazione

Esercizio: sfruttando la linearità dell'aspettazione, dimostrare che:

$\sigma_A^2 = M_A - m_A^2$

Se conosciamo solo media e SD ma NON la ddp, si possono valutare solo dei bounds per gli intervalli di confidenza:

Disuguaglianza di Chebychev

$$Prob[|a-m|>kSD] \leq 1/k^2 \text{ ovvero } Prob[|a-m|<kSD] \geq 1-1/k^2$$

K è un qualsiasi numero positivo : al crescere di k, la prob che a si discosti dal suo valor medio per una quantità via via più grande è via via più piccola

K=1 $Prob[m-SD<a<m+SD] \geq 0$

K=2 $Prob[m-2SD<a<m+2SD] \geq 1-1/4=75\%$

K=3 $Prob[m-3SD<a<m+3SD] \geq 1-1/9=89\%$

K=4 $Prob[m-4SD<a<m+4SD] \geq 1-1/16=94\%$

almeno il 75% dei valori della v.a. cadono nell'intervallo media-2SD, media+2SD

almeno l'89% dei valori della v.a. cadono nell'intervallo media-3SD, media+3SD

almeno il 94% dei valori della v.a. cadono nell'intervallo media-4SD, media+4SD

Esempio: età di un individuo di nazionalità italiana

Non conosciamo la ddp, ma solo

Media = 42

SD = 22

In realtà ben 98.3% degli italiani ha un'età minore di 86 anni, quindi l'informazione che si deriva dalla disuguaglianza di chebychev è corretta ma non è molto precisa

Possiamo dire

K=2 Prob[0<a<86]≥=75%

almeno il 75% degli italiani ha una età compresa tra 0 e 86 anni

Esempio: glicemia a digiuno della popolazione normale

Non conosciamo la ddp, ma solo

Media = 80

SD = 8

Possiamo dire

K=2 Prob[64<a<96]≥=75%

almeno il 75% dei soggetti normali ha una glicemia compresa tra 64 e 96mg/100ml

K=3 Prob[56<a<104]≥=89%

almeno l'89% dei soggetti normali ha una glicemia compresa tra 56 e 104mg/100ml

Modelli di distribuzione di v.a.

Discreta :

- binomiale

Continua:

- gaussiana o normale
- log normale
- uniforme

v.a. binomiale (di Bernoulli)

Partiamo da una v.a. discreta, che può assumere solo due stati (es. 0 e 1 oppure testa e croce). Supponiamo che la prob di occorrenza del 1° stato sia p. Quella del secondo sarà (1-p). Quindi:

A = [0 1] P[0]=p P[1]=1- p

Supponiamo ora di condurre N esperimenti indipendenti (es lancio la moneta N volte). Tale esperimento si chiama processo di Bernoulli. L'esito di ogni prova è descritto da una v.a. A e queste sono tra loro indipendenti. Indichiamo con X la variabile aleatoria che conta il numero di volte che l'esito è pari a zero (testa). X è una v.a. discreta binomiale con la seguente ddp

$$\mathrm{Prob}[X=k]=\binom{N}{k}p^{k}(1-p)^{N-k}$$

$$\mathrm{media}=Np$$

$$SD=\sqrt{Np(1-p)}$$

Esempio 1

Lancio di una moneta 2 esiti: T e C, ognuno con probabilità 0.5

Considero 4 lanci

Definisco x = numero di volte in cui l'esito è T

$$\text{Prob}[X=k]=\binom{4}{k}0.5^{k}0.5^{4-k}=\binom{4}{k}0.5^{4}=\binom{4}{k}0.0625$$

$$\text{Prob}[X=0]=\binom{4}{0}0.0625=1\cdot 0.0625=0.0625$$

$$\text{Prob}[X=1]=\binom{4}{1}0.0625=4\cdot 0.0625=0.25$$

$$\text{Prob}[X=2]=\binom{4}{2}0.0625=\frac{4\cdot 3}{2}\cdot 0.0625=0.375$$

$$\text{Prob}[X=3]=\binom{4}{3}0.0625=\frac{4\cdot 3\cdot 2}{3\cdot 2}\cdot 0.0625=0.25$$

$$\text{Prob}[X=4]=\binom{4}{4}0.0625=1\cdot 0.0625=0.0625$$

Esercizio 1: Verificare

$$\sum_{k=0}^{4}\text{Prob}[X=k]=1$$

$$\text{media}=Np=2$$

$$SD=\sqrt{Np(1-p)}=1$$

Esercizio 2: ripetere l'esempio per una moneta truccata, con Prob[T]=0.7, Prob[C]=0.3

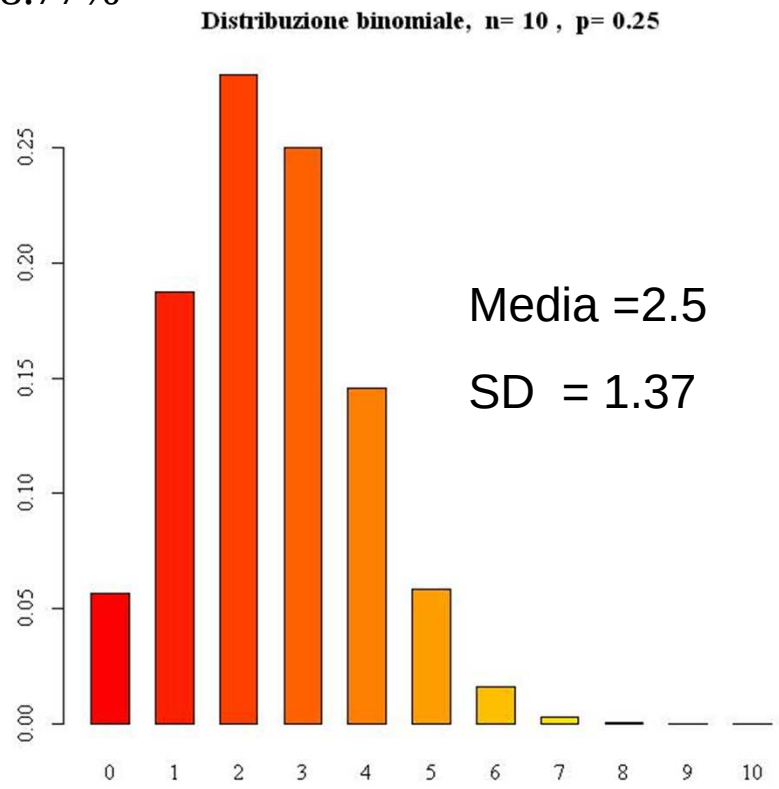
Esempio 2

Un classificatore automatico ha una probabilità di commettere un errore pari a 0.25.

Viene applicato a 10 soggetti.

Qual è la probabilità che uno dei 10 i soggetti sia classificato erroneamente ?

$$\text{Prob}[X=1]=\binom{10}{1}0.25^{1}\cdot 0.75^{9}=10\cdot 1.877=18.77\%$$



v.a. gaussiana o normale

Riveste un ruolo importante nella teoria della probabilità

Molte grandezze antropometriche/fisiologiche come statura, peso, pressione arteriosa, glicemia in prima approssimazione possono essere descritte con una legge gaussiana

Si manifesta quando la v.a. osservata è il risultato della somma di un numero sufficientemente grande di variabili aleatorie indipendenti (o al limite debolmente indipendenti) che obbediscono a leggi di distribuzioni diverse

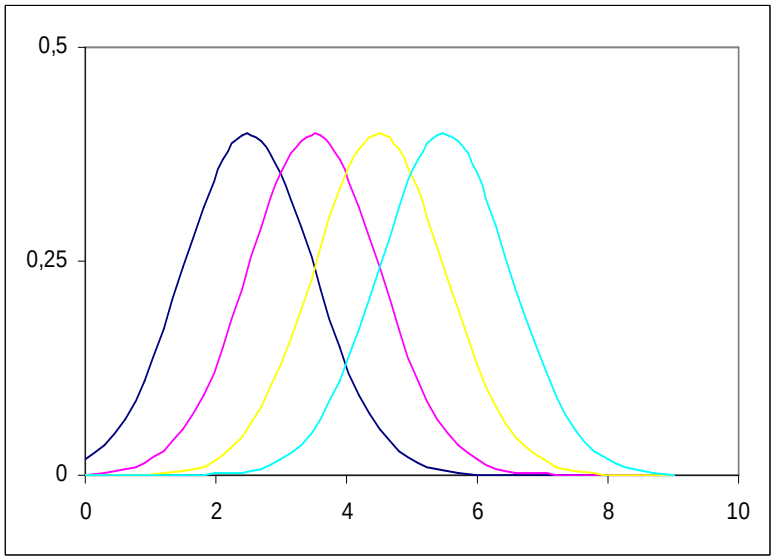
Numerose variabili aleatorie di uso comune, quali ad es. gli errori di misura, si possono rappresentare come somma di singoli termini, ciascuno dei quali dovuto ad una causa non dipendente dalle altre. Se ognuno di tali termini influisce relativamente poco sulla somma, quest'ultima può essere approssimata con una legge normale

v.a. gaussiana

media= m

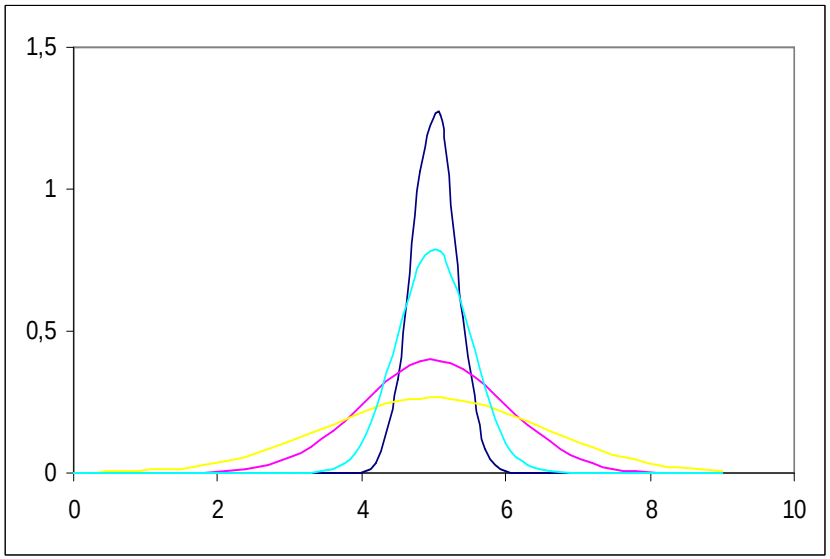
SD = σ

$$f_A(a) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{a-m}{\sigma}\right)^2}$$



Media diversa

SD uguale



Media uguale

SD diversa

v.a. gaussiana normalizzata: Media=0 SD=1

Esistono tabelle con i valori della distribuzione

$$g(a) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a^2}$$

a	G(a)=P[A≤a]
-3.0	1.3498980e-03
-2.5	6.2096653e-03
-2.0	2.2750132e-02
-1.5	6.6807201e-02
-1.0	1.5865525e-01
-0.5	3.0853754e-01
0.0	5.0000000e-01
0.5	6.9146246e-01
1.0	8.4134475e-01
1.5	9.3319280e-01
2.0	9.7724987e-01
2.5	9.9379033e-01
3.0	9.9865010e-01

Data una variabile gaussiana generica con

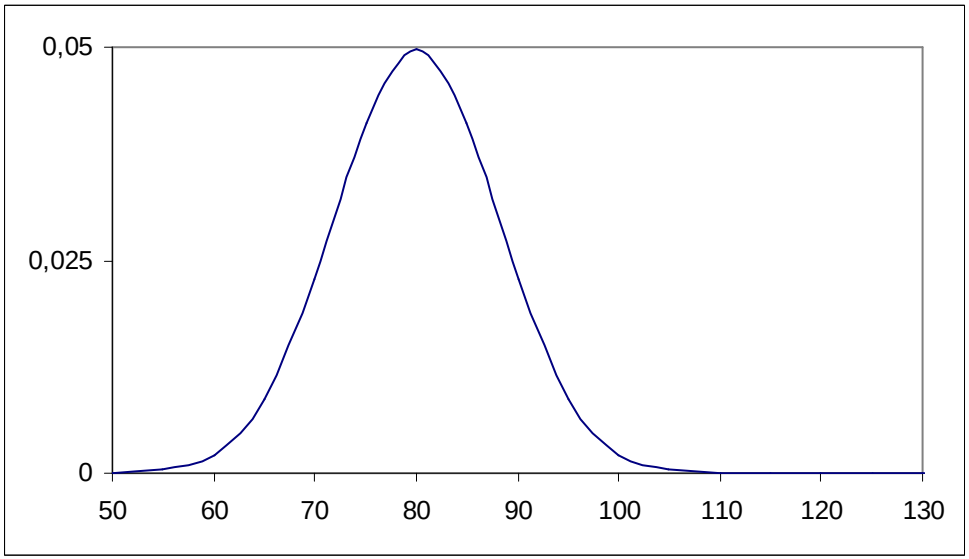
- media=m
- SD=σ

ci si riporta alla gaussiana normalizzata con la trasformazione:

(a-m)/σ

Esempio

La glicemia a digiuno della popolazione normale è una v.a.gaussiana con media=80 (mg/dl) e SD=8



$$f_A(a) = \frac{1}{\sqrt{2\pi}\cdot 8} e^{-\frac{1}{2}\left[\frac{a-80}{8}\right]^2}$$

Un soggetto ha la glicemia a digiuno pari a 104 mg/dl.
Può essere normale?
In teoria si, perché in base al modello assunto, tutti i valori della glicemia sono ammissibili
Possiamo però valutare la verosimiglianza di tale valore, analizzando la probabilità di avere un valore maggiore di 104

Esempio

Normalizzazione (104-80)/8=3

a	G(a)=P[A≤a]
-3.0	1.3498980e-03
-2.5	6.2096653e-03
-2.0	2.2750132e-02
-1.5	6.6807201e-02
-1.0	1.5865525e-01
-0.5	3.0853754e-01
0.0	5.0000000e-01
0.5	6.9146246e-01
1.0	8.4134475e-01
1.5	9.3319280e-01
2.0	9.7724987e-01
2.5	9.9379033e-01
3.0	9.9865010e-01

G(3)=99.865%

Solo 0.15% dei soggetti hanno una glicemia maggiore di 104! Quindi il valore 104 ha una probabilità molto bassa di essere il valore di un soggetto normale!

v.a. gaussiana: intervalli di confidenza

Sono gli intervalli di valori in cui, con una certa probabilità, cadono i valori della v.a

Per una variabile gaussiana, descritta univocamente da media e SD, questi dipendono solo dalla media e dalla SD

$P[m-SD \leq A \leq m+SD] =$

$F_A(m+SD)-F_A(m-SD)=G(1)-G(-1)=$

$0.84-0.16=0.68$

→ ***68% dei valori della v.a. sono compresi tra m-SD e m+SD***

$P[m-2SD \leq A \leq m+2SD] = 95\%$

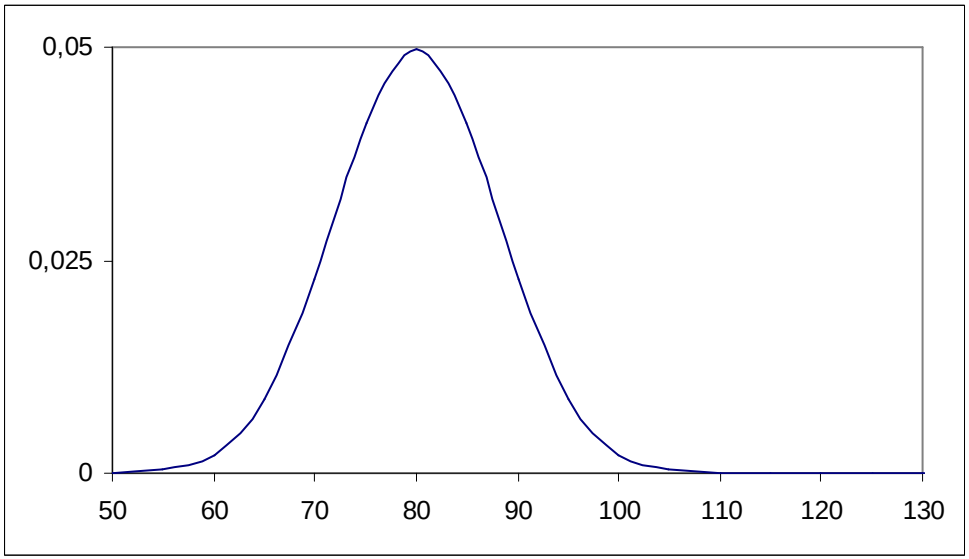
→ ***95% dei valori della v.a. sono compresi tra m-2SD e m+3SD***

$P[m-3SD \leq A \leq m+3SD] = 99.7\%$

→ ***Praticamente tutti i valori (ameno del 3 per mille) sono compresi tra m-3SD e m+3SD***

Esempio

La glicemia a digiuno della popolazione normale è una v.a.gaussiana con media 80 (mg/dl) e SD 8



$$f_A(a) = \frac{1}{\sqrt{2\pi}\cdot 8} e^{-\frac{1}{2}\left[\frac{a-80}{8}\right]^2}$$

P[m-SD ≤ A ≤ m+SD] = P[72≤ A ≤ 88] = **68%**

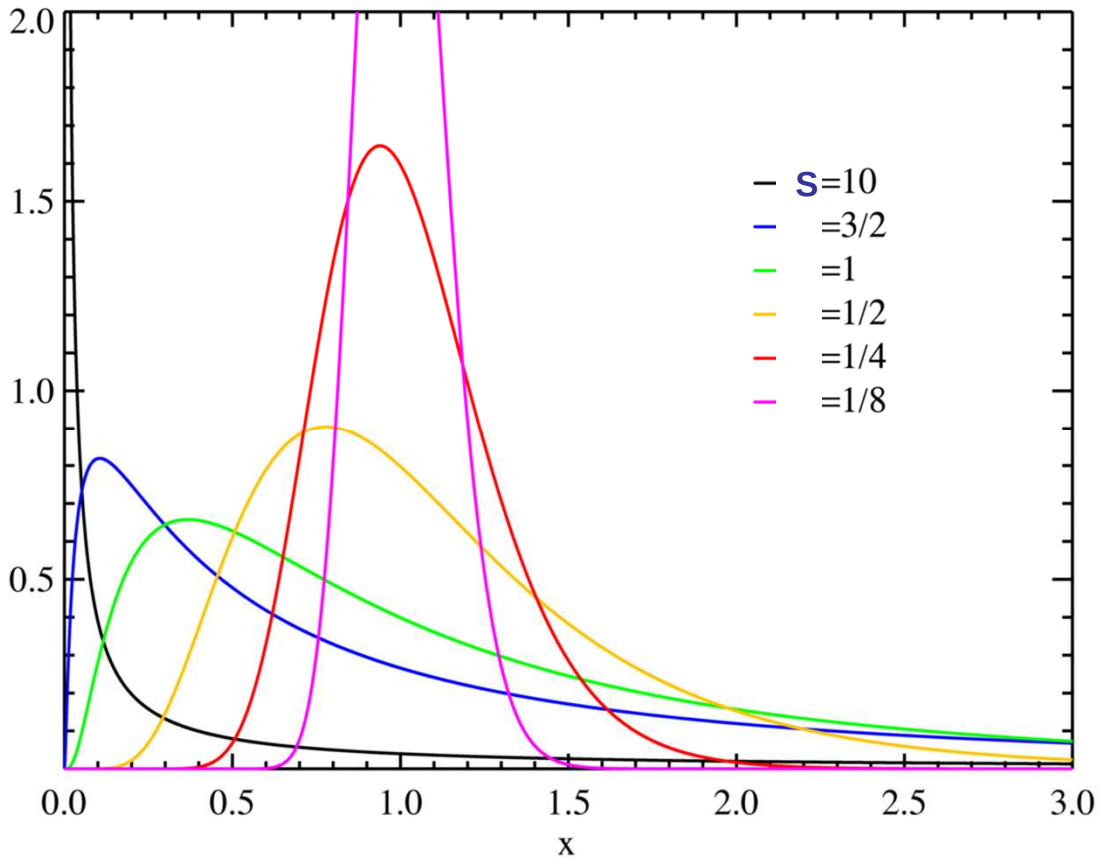
P[m-2SD ≤ A ≤ m+2SD] = P[64≤ A ≤ 96] = **95%**

P[m-3SD ≤ A ≤ m+3SD] = P[56≤ A ≤ 104] = **99.7%**

Intervalli di confidenza

Se confrontiamo con i valori forniti dalla dis di Chebychev, è evidente che conoscere la ddp permette di avere informazioni più precise sulla v.a.

v.a. log-normale



$$f_A(a) = \frac{1}{a\sqrt{2\pi}s} e^{-\frac{1}{2}\left(\frac{\log(a)-m}{s}\right)^2}$$

log(a) ha distribuzione normale

Sono ammissibili solo valori positivi di a, quindi questo è un modello molto usato per descrivere variabili fisiologiche che possono assumere valori solo positivi

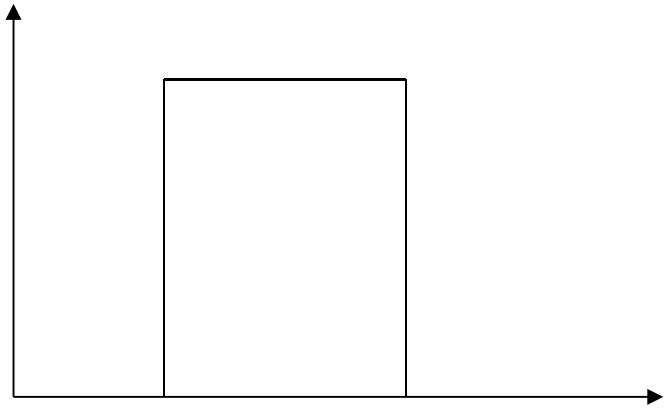
$$\text{media} = e^{m+s^2/2}$$

$$SD = e^{2(m+s^2)} - e^{2m+s^2}$$

v.a. uniforme

$f_A(a)=1/(a_{max}-a_{min})$ $a_{min}<a<a_{max}$

$= 0$ altrove



Esercizio:

Verificare che

$$\text{media} = \frac{a_{min} + a_{max}}{2}$$

$$SD = \frac{a_{max} - a_{min}}{\sqrt{12}}$$

Riassumendo

- v.a. : è uno strumento comodo per descrivere fenomeni casuali
- Una v.a. è caratterizzata completamente dalla sua densità (o distribuzione) di probabilità
- Note densità o distribuzione, si possono valutare gli intervalli di confidenza
- Media e SD danno una caratterizzazione sintetica della v.a.
- Note media e SD, ma non densità o distribuzione, si possono valutare solo dei bounds per gli intervalli di confidenza
- Le v.a. gaussiane sono completamente caratterizzate da media e SD

Generazione di v.a

Normale

MATLAB
x=randn(100,1)

Genera 100 valori di una v.a. gaussiana normalizzata

MATLAB
x=randn(100,1)
y=m+s*x

Genera 100 valori di una v.a. gaussiana media m, SD=s

Uniforme

MATLAB
x=rand(100,1)

Genera 100 valori di una v.a. uniforme in [0,1]

MATLAB
x=rand(100,1)
y=a+(b-a)*x

Genera 100 valori di una v.a. uniforme in [a,b]

Variabili aleatorie vettoriali

Il concetto di v.a. si può estendere al caso di due (o più) dimensioni, sia per v.a. discrete che continue, se ad ogni evento elementare associamo due o più funzioni.

Esempio:

Si consideri il lancio di due dadi e si associ ad ognuno dei 36 possibili eventi elementari (equiprobabili) una coppia di v.a, la prima (A) pari alla differenza in valore assoluto dei numeri che appaiono sulle due facce e la seconda (B) alla loro somma:

A=[0,1,2,3,4,5]

B=[2,3,4,5,6,7,8,9,10,11,12]

P[A=a_i, B=b_j] è la somma delle probabilità degli eventi elementari per i quali differenza=a_i, somma=bj, da cui per es.

P[0,2]=1/36 perché solo l'evento el (1,1) ha diff=0, somma=2

P[2,2]=0 perché nessun evento el ha diff=2, somma=2

P[2,4]=2/36 perché gli eventi el (3,1) e (1,3) hanno diff=2, somma=4

Esempio:

Si consideri una persona di nazionalità italiana e si consideri la v.a. bidimensionale che ha come prima componente il peso e come seconda la sua altezza

Variabili aleatorie vettoriali-dim2

Estendendo le definizioni già viste per il caso monodimensionale, si possono definire per v.a. vettoriali di dim 2, nel caso discreto le P bidimensionali

P[A=ai, B=bj]

nel caso continuo le funzioni distribuzione e ddp congiunte (bidimensionali)

$$F_{AB}(a,b) = \text{prob}[A < a, B < b]$$

$$f_{AB}(a,b) = \frac{\partial^2 F_{AB}(a,b)}{\partial a \partial b}$$

Indipendenza

Dato un vettore di due variabili aleatorie continue A e B queste sono indipendenti se $f_{AB}(a,b)=f_A(a)f_B(b)$. Analogamente nel caso discreto

Esempio:

Nell'esempio precedente del lancio di due dadi, le due v.a. A e B NON sono indipendenti. Infatti, ad es. P[A=0]=6/36; P[B=2]=1/36 e P[A=0,B=2]=1/36≠P[0]P[2]

La stessa conclusione ragionevolmente varrà per peso e altezza di un soggetto, che è verosimile pensare NON siano indipendenti tra loro

Alcuni elementi di statistica inferenziale

Stima: scoprire la distribuzione di una popolazione, a partire da informazioni contenute in un campione casuale estratto da essa

Test di ipotesi: utilizzare la descrizione statistica per prendere delle decisioni

Stima di una ddp

Consideriamo due alternative:

- È noto (o si ipotizza) il modello probabilistico, es. si tratta di una ddp gaussiana oppure uniformemente distribuita oppure, nel caso discreto, binomiale. Il problema di stima è allora riportato ad un problema di stima dei parametri del modello, cioè m e s nel caso gaussiano, a_{min} e a_{max} nel caso di v.a. uniformemente distribuita, p nel caso binomiale.
- Non si conosce il modello probabilistico, ovvero non si vogliono fare ipotesi su di esso.

  Consideriamo dapprima la prima situazione, e supponiamo di voler stimare i parametri media m e varianza s^2 che caratterizzano completamente una v.a. gaussiana. Gli stimatori di massima verosimiglianza di media m e varianza s^2 coincidono con media e varianza campionaria

Stimatore della media

$$T = \frac{1}{N}\sum_{i=1}^{N} A_i$$

Esercizio: Verificare che lo stimatore è

non polarizzato

La varianza di quasto stimatore è pari a s^2/N dove s^2 è la varianza vera ma incognita della v.a. Quindi lo stimatore è anche consistente

Stimatore della varianza

1° stimatore

$$T = \frac{1}{N}\sum_{i=1}^{N}(A_i - A_m)^2$$

E' uno stimatore

polarizzato e consistente

2° stimatore

$$T = \frac{1}{N-1}\sum_{i=1}^{N}(A_i - A_m)^2$$

E' uno stimatore

Non polarizzato e consistente

Una volta che si è eseguito il campionamento, e quindi sono disponibili i dati, i loro valori vengono inseriti negli stimatori, ottenendo così le stime di media e varianza

Stima della media

$$\hat{m} = \frac{1}{N}\sum_{i=1}^{N} a_i$$

Stima della varianza

1° stimatore

$$\hat{var} = \frac{1}{N}\sum_{i=1}^{N}(a_i - \hat{m})^2$$

2° stimatore

$$\hat{var}' = \frac{1}{N-1}\sum_{i=1}^{N}(a_i - \hat{m})^2$$

Se la v.a. non è gaussiana?

Bisogna di volta in volta definire degli stimatori dei parametri della ddp, ad es utilizzando la max verosimiglianza

Comunque con gli stimatori definiti in precedenza possiamo sempre stimare media e varianza della ddp. In alcuni casi è possibile risalire da media e varianza ai parametri della distribuzione

Esercizio

Assunto un modello di ddp uniformemente distribuita, ricavare i suoi parametri (limiti superiore ed inferiore dell'intervallo in cui la ddp è non nulla) dalle stime di media e varianza ottenute da un campione estratto dalla popolazione

Riassumendo

- Essendo basato su v.a., lo stimatore è anch'esso una v.a.
- Criteri fondamentali per valutare la bontà di uno stimatore: non polarizzazione e varianza «piccola»
- E' utile fornire insieme alla stima un indice della sua affidabilità, ad es. dando la sua varianza (se è piccola, la stima è precisa)

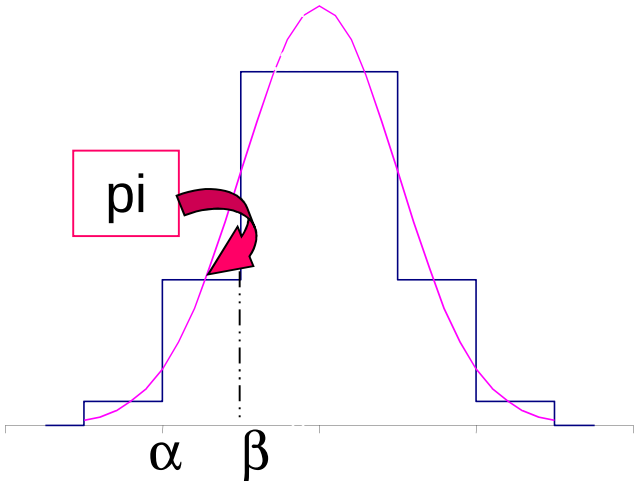
Stima della densità di probabilità:

Consideriamo ora il problema di stimare la ddp di una v.a. a partire da un insieme di suoi valori. Il problema è più complesso, perché si tratta di stimare non più un parametro (come media e SD) ma una funzione. Esistono vari approcci.

Metodo dell'istogramma

Si approssima la densità di probabilità da stimare con una funzione continua a tratti. Preso un generico intervallo, di estremi α e β, indichiamo con pi il valore dell'approssimazione nell'intervallo.
Per calcolare pi:

$$\int_{\alpha}^{\beta} f_A(a)da = \mathrm{Prob}(a \in [\alpha,\beta]) = pi * (\beta - \alpha)$$



Quindi:

$$pi = \frac{\mathrm{Prob}(a \in [\alpha,\beta])}{(\beta - \alpha)}$$

stima della $\mathrm{Prob}(a \in [\alpha,\beta])$

Indichiamo con $P_{\alpha\beta}$ tale probabilità. Essa rappresenta la probabilità che un valore della variabile aleatoria preso a caso appartenga all'intervallo $[\alpha,\beta]$. Consideriamo ora gli N dati che abbiamo a disposizione. Chiamiamo X la variabile aleatoria che conta il numero di volte che un elemento cade nell'intervallo $[\alpha,\beta]$. Sappiamo che se gli N dati sono indipendenti, X ha una distribuzione binomiale:

$$\mathrm{Prob}[X=k]=\binom{N}{k}P_{\alpha\beta}{}^{k}(1-P_{\alpha\beta})^{N-k}$$

con $E[X]=NP_{\alpha\beta}$ ovvero $P_{\alpha\beta}=E[X]/N$. Quindi per misurare la $P_{\alpha\beta}$ dovremmo conoscere il valor medio della variabile X, cioè il valor medio del numero di elementi che cadono nell'intervallo. Noi però, contando quanti elementi del nostro insieme di N dati cadono nell'intervallo (supponiamo siano k) abbiamo a disposizione solo una realizzazione di tale variabile, e NON il suo valor medio. Questo ci fornirà una stima della probabilità cercata:

$$\hat{P}_{\alpha\beta}=k/N$$

Ricapitolando:

$$pi = \frac{Prob(a \in [\alpha,\beta])}{\beta - \alpha} = \frac{P_{\alpha\beta}}{\beta - \alpha}$$

$$\hat{P}_{\alpha\beta} = k / N$$

Numero di dati che cadono nell'intervallo α,β

Numero di dati a disposizione

$$\hat{p}i = \frac{k / N}{(\beta - \alpha)}$$

Stima della ddp nell'intervallo α,β

Se confrontiamo con la definizione di istogrammma data in precedenza, la ddp viene stimata dividendo l'istogramma per il numero totale di campioni e l'ampiezza dell'intervallo

Verificare che la ddp così stimata soddisfa sempre la proprietà che l'area sottesa dalla ddp è pari ad uno

Alcune osservazioni:

E' critica la scelta dell'ampiezza degli intervalli:

Per avere una buona approssimazione della ddp con una funzione continua a tratti → Intervalli piccoli

Per avere una buona stima della E[k] → K elevato, quindi intervalli grandi

Bisogna scegliere gli intervalli in modo da mediare tra queste due esigenze. Comunque si avrà sempre un certo grado di approssimazione e un certo errore di stima

3° problema: test di ipotesi

Lo scopo è aiutare a prendere una decisione su qualche caratteristica di una (o più) popolazioni esaminando l'andamento di tale caratteristica in campioni estratti dalle popolazioni

Esempio

La glicemia a digiuno in una popolazione normale dipende dall'eta?

Misuro la glicemia a digiuno:

- 15 soggetti giovani (18-30 anni)
- 15 soggetti anziani (più di 65 anni)

Sulla base di questi valori, devo decidere se la glicemia a digiuno è uguale o diversa in queste due popolazioni

Passi fondamentali di un test di ipotesi

1. Analisi dei dati disponibili

   Ogni test lavora su un certo tipo di dati, es. qualitativi o quantitativi, discreti o continui

2. Verifica degli assunti di base

   Ogni test fa delle assunzioni. Queste vanno verificate sui dati prima di applicare il test

3. Impostazione dell'ipotesi statistica

   Si formula l'ipotesi che si vuol testare (ipotesi nulla)

4. Costruzione della statistica

   Viene utilizzata per prendere delle decisioni riguardo all'ipotesi formulata

5. Determinazione della distribuzione della statistica

   Se l'ipotesi è corretta, la statistica ha una certa distribuzione

6. Definizione della regola di decisione

   Sono le condizioni per cui si accetta o si rifiuta l'ipotesi nulla, sulla base di una probabilità di errore

Test di Student

Serve per testare se due campioni provengono dalla stessa popolazione, oppure da popolazioni diverse.

1. Analisi dei dati disponibili

I dati sono di tipo quantitativo e misurabili su una scala continua

2. Verifica degli assunti di base

I dati costituiscono due campioni indipendenti, di dimensioni N_1 e N_2, estratti da popolazioni gaussiane con medie (m_1 e m_2) e varianze note (s_1^2 e s_2^2)

3. Impostazione dell'ipotesi statistica

Ipotesi da sottoporre a verifica (ipotesi nulla)

H_0: m1 = m2

Ipotesi alternativa

H_1: m1 ≠ m2

4. Costruzione della statistica

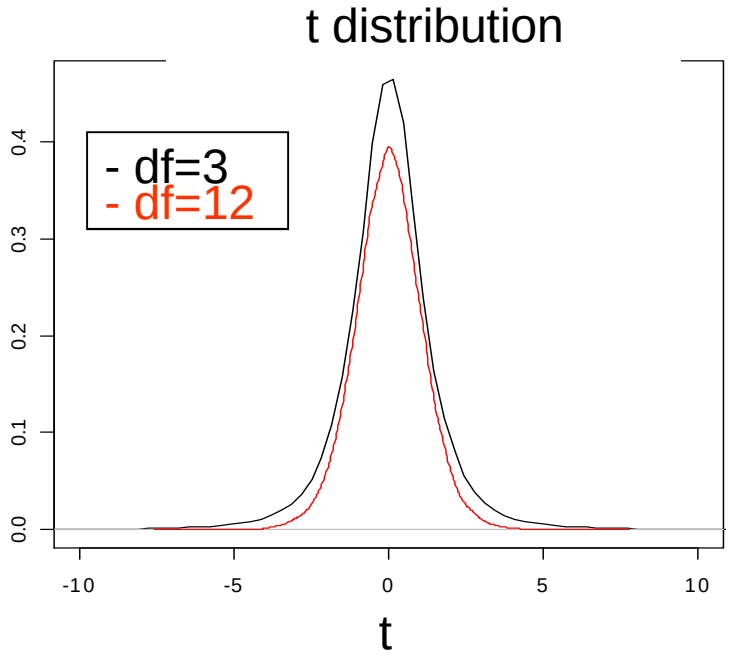
$$t = \frac{m_1 - m_2}{s_p \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

dove s_p è una media ponderale di s_1 e s_2

$$s_p^{\,2} = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}$$

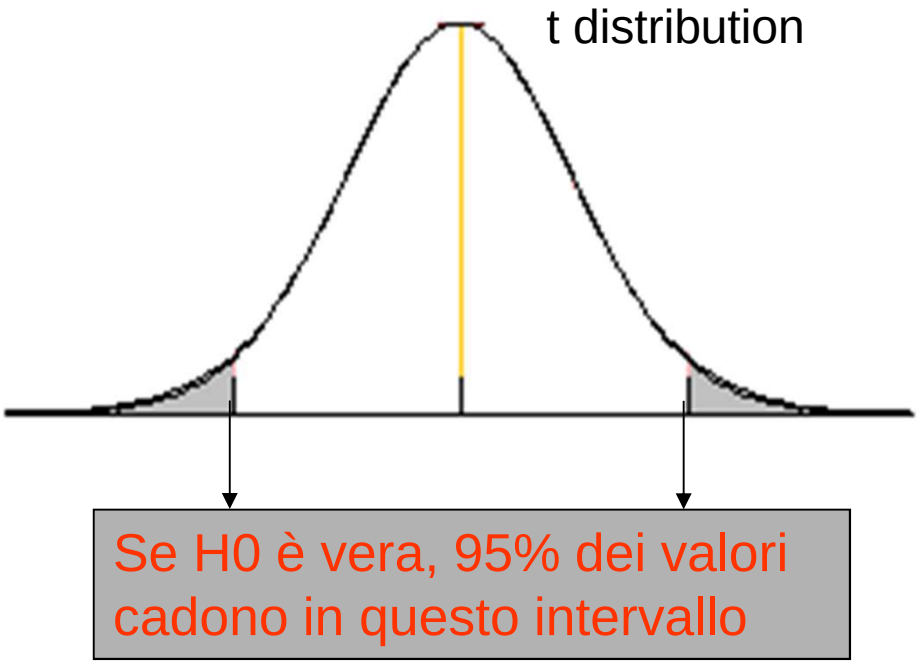
5. Determinazione della distribuzione della statistica

Sotto H_0:
la v.a. t ha una distribuzione nota, detta distribuzione di Student, la cui forma dipende unicamente dai gradi di libertà $df=N_1+N_2-2$



6. Definizione della regola di decisione

Sotto H_0:
Fissato un livello di significatività α (es. α=5%) valuto l'intervallo di confidenza per la v.a. t



Regola di decisione:

Accetto l'ipotesi H_0 se t cade all'interno dell'intervallo di confidenza

Rifiuto l'ipotesi H_0 se t cade al di fuori dell'intervallo di confidenza, perché sono valori poco probabili

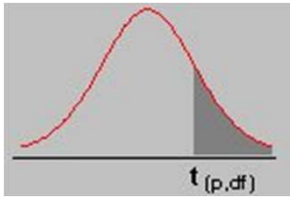
Esempio

S_p= 8.51
t= 2.57
Df= 28

Glicemia a digiuno:

- 15 soggetti giovani m_1= 80 σ_1= 8
- 15 soggetti anziani m_2= 88 σ_2= 9

t table with right tail probabilities



df\p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
5	0,2672	0,7267	1,4759	2,0150	2,5706	3,3649	4,0321	6,8688
6	0.264835	0.717558	1,439756	1,943180	2,44691	3,14267	3,70743	5,9588
7	0.263167	0.711142	1,414924	1,894579	2,36462	2,99795	3,49948	5,4079
8	0.261921	0.706387	1,396815	1,859548	2,30600	2,89646	3,35539	5,0413
9	0.260955	0.702722	1,383029	1,833113	2,26216	2,82144	3,24984	4,7809
10	0.260185	0.699812	1,372184	1,812461	2,22814	2,76377	3,16927	4,5869
28	0.255768	0.683353	1,3125	1,7011	2,0484	2,4671	2,7633	3,6739

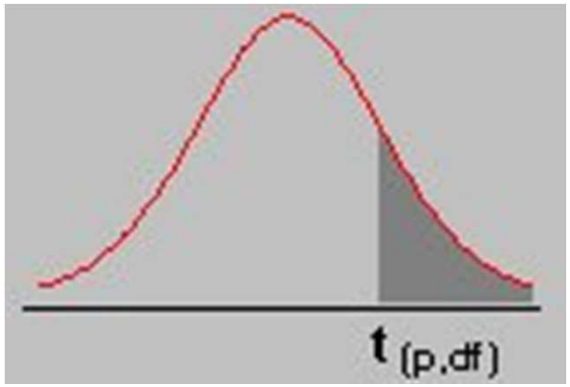
Regola di decisione:

Il valore di t (nell'es pari a 2,57) cade al di fuori dell'intervallo di confidenza (-/+2,0484)? SI, quindi rifiuto H_0 ovvero concludo che le due medie sono diverse

In alternativa, con Matlab:

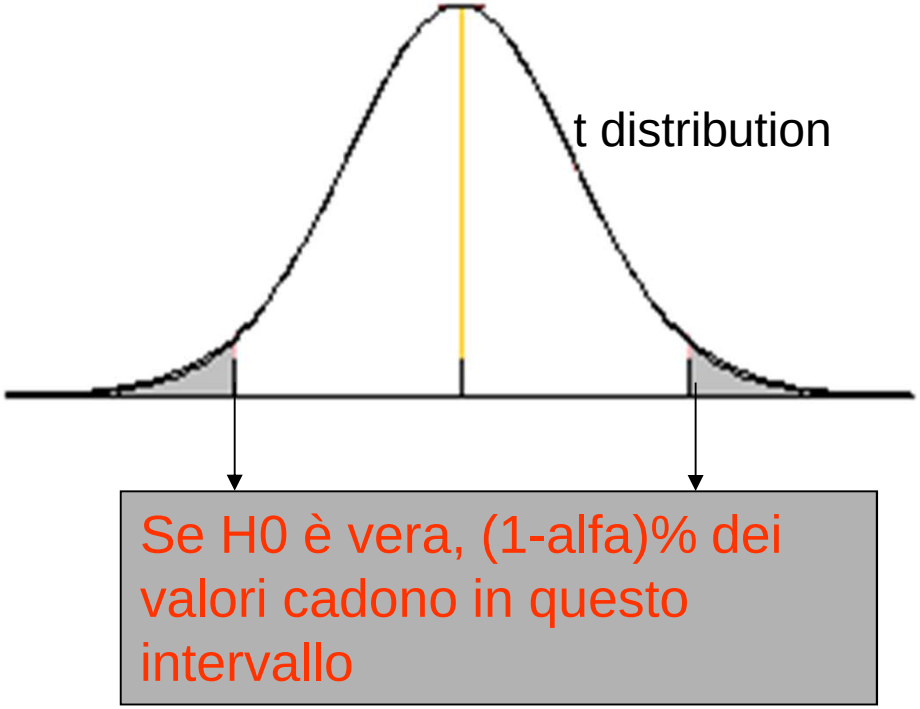
MATLAB
p=tcdf(t,df)

Calcola l'integrale da meno infinito a t



Il complemento a 1 di tcdf dà la probabilità che, sotto Ho, la variabile di Student sia maggiore del valore trovato t. Questa probabilità viene spesso chiamata p-value. Se assume valori bassi, es. <0.025, si rifiuta Ho, altrimenti si accetta

Sul significato di alfa



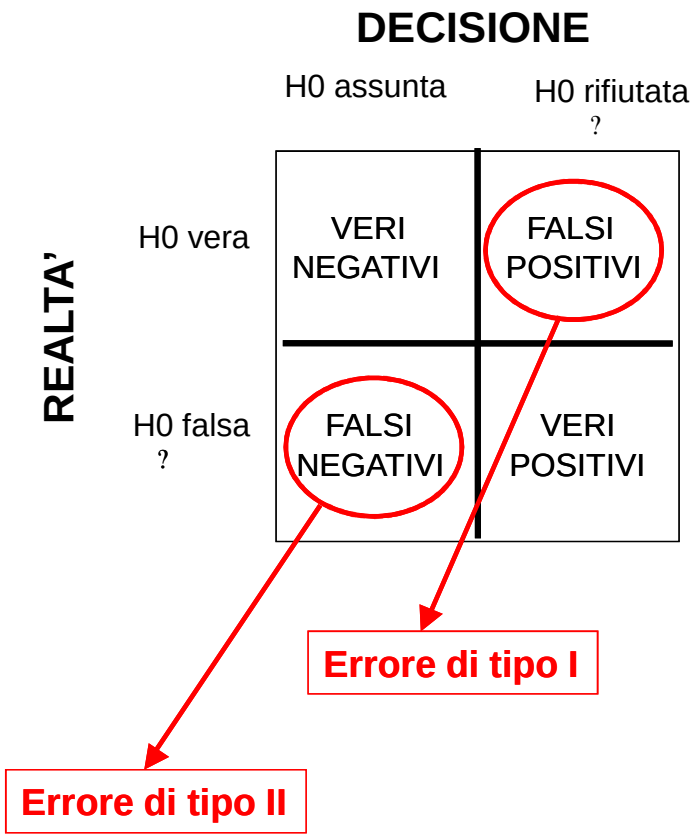
Regola di decisione:

Accetto l'ipotesi H_0 se t cade all'interno dell'intervallo di confidenza

Rifiuto l'ipotesi H_0 se t cade al di fuori dell'intervallo di confidenza, anche se in realtà, con H_0 vera, ho una probabilità 5% che i valori di t cadano fuori dell'intervallo. Quindi 5% è la probabilità di commettere un errore

Quale tipo di errore?

Errori associati ad un test



Nomenclatura:

Negativo/positivo è legato all'esito del test:

negativo se assumo H_0, perché vuol dire che il test non trova differenze

positivo se rifiuto H_0, perché vuol dire che il test trova una differenza

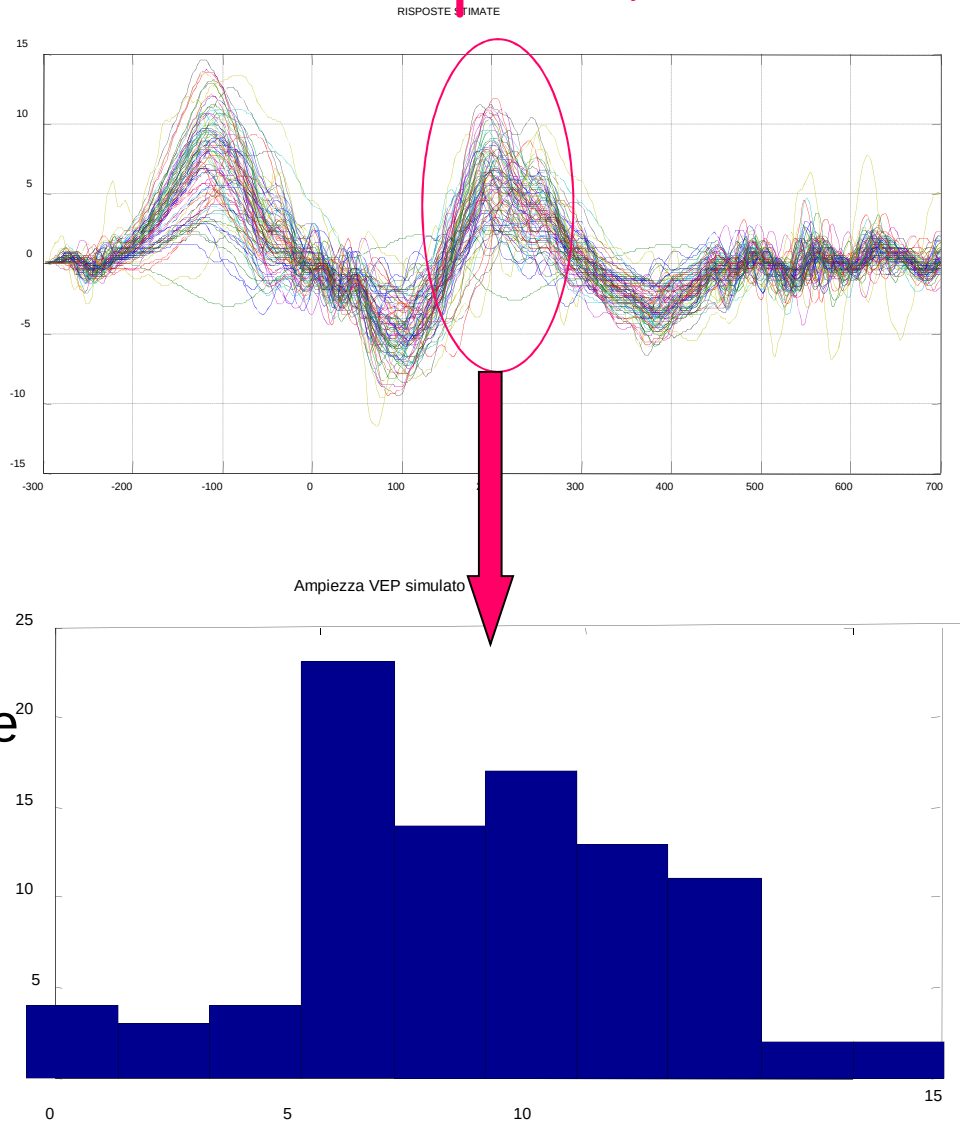
Vero /falso è legato alla correttezza del risultato

Alfa quantifica l'errore di tipo I, cioè la probabilità di rifiutare l'ipotesi H_0 quando questa invece è corretta

Nulla si può dire riguardo all'errore di tipo II

Questi strumenti sono molto utili nell'analisi di segnali biologici, ad es. per rappresentare la variabilità di alcuni parametri

Interessa il parametro VEP=ampiezza potenziale visivo. Lo caratterizziamo con media, SD e istogramma

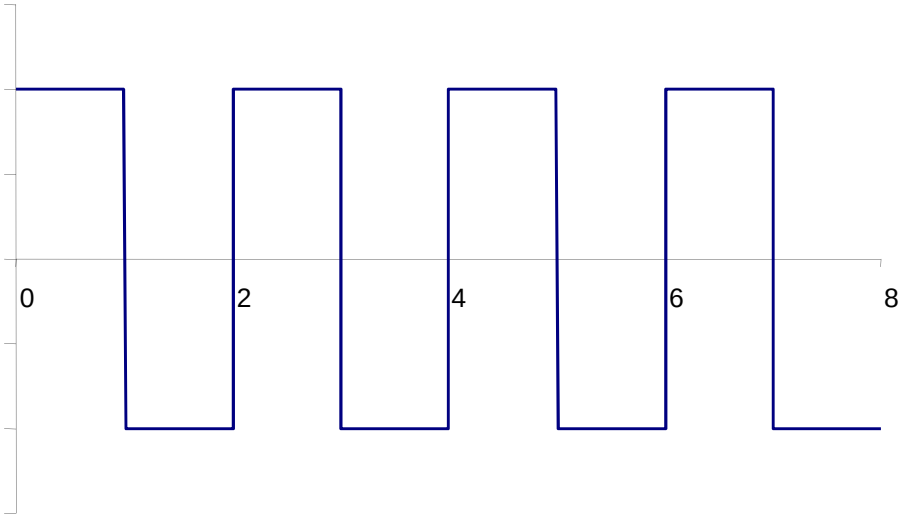


Potenziali evocati cognitivi misurati in un paziente in seguito all'applicazione ripetuta dello stesso stimolo.

MEDIA=7.3
SD=2.4

…oppure per rappresentare la variabilità dei valori assunti da un segnale…..

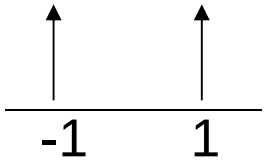
Esempio



v.a. discreta

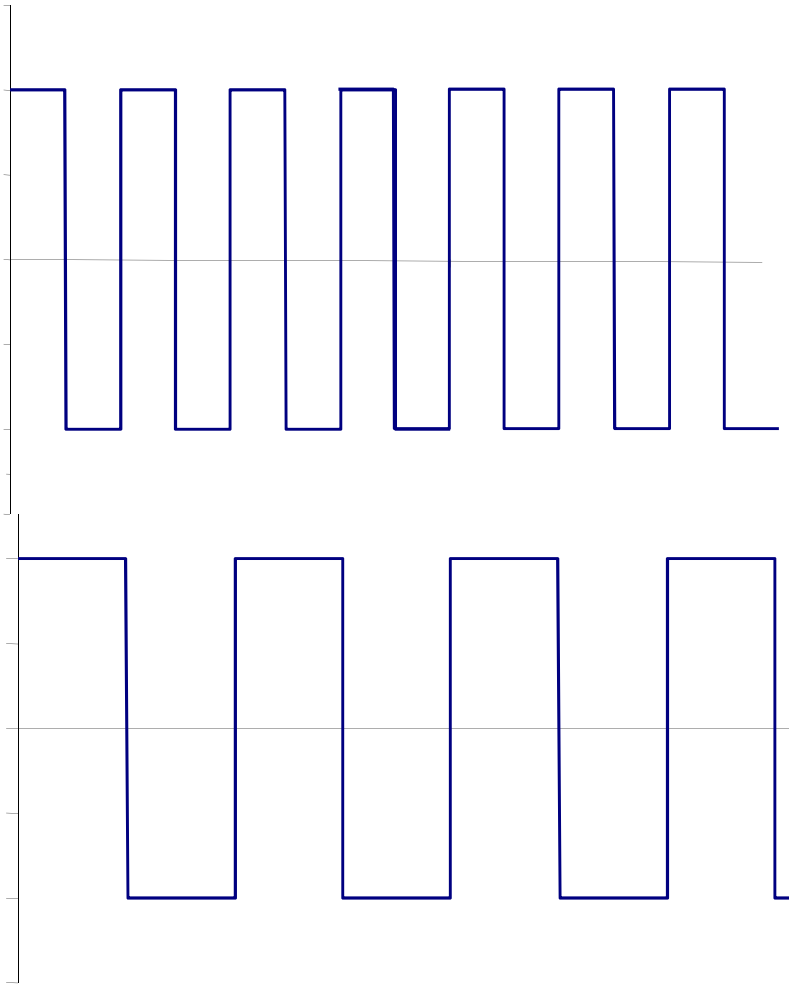
Media=0

SD=1



….. ma NON identifica in modo univoco il segnale, perché non tiene conto di come i valori sono legati tra loro nel tempo

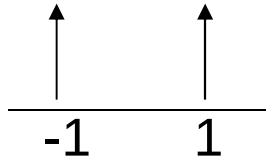
Infatti questi due segnali hanno la stessa ddp



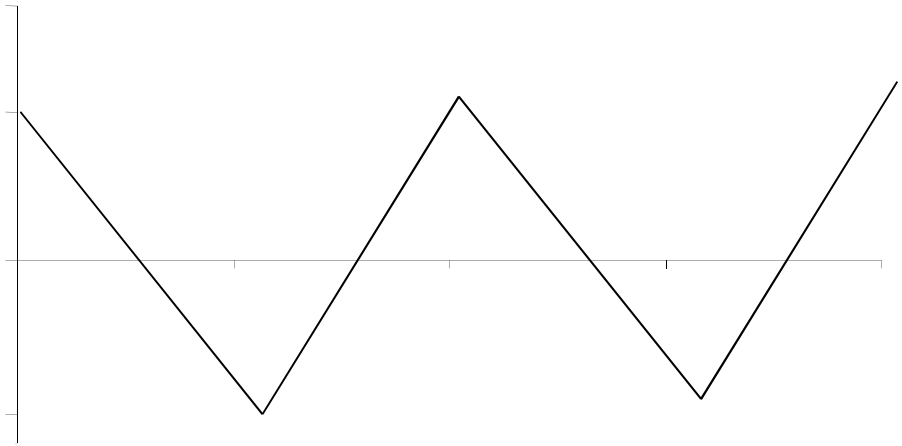
v.a. discreta

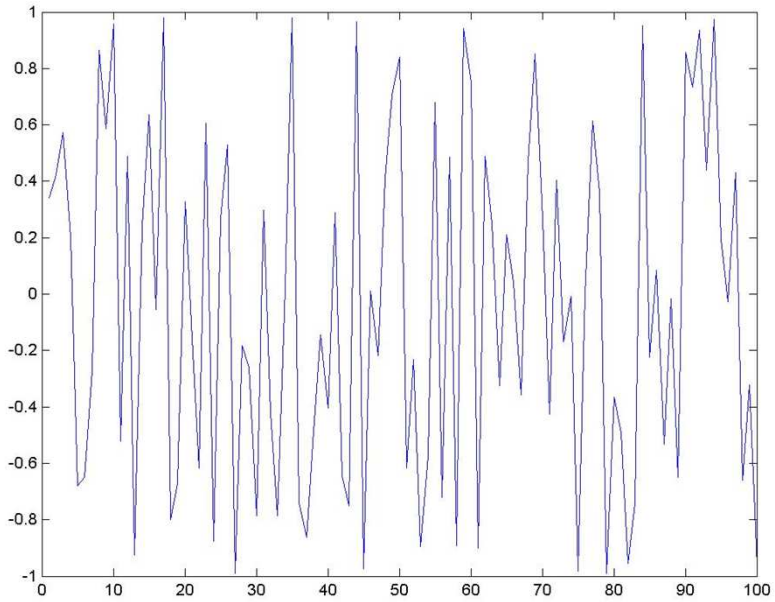
Media=0

SD=1



Anche questi due segnali hanno la stessa ddp



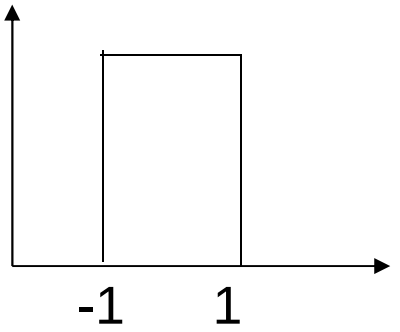


v.a. discreta compresa tra -1 e 1

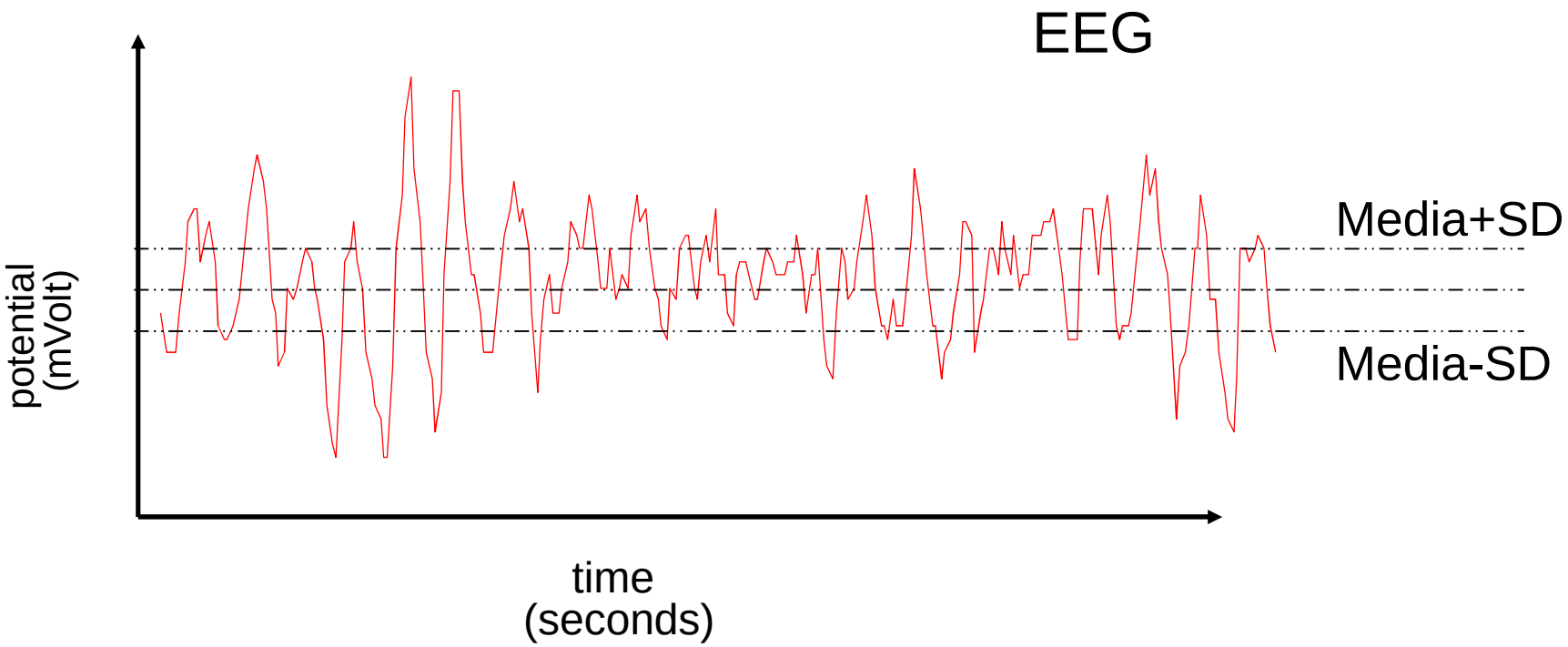
uniformemente distribuita

Media=0

SD=4/rad(12)



Media, varianza e ddp danno informazioni sui valori assunti da un segnale, ma NON sulla sua forma



Serve una descrizione statistica di grandezze che evolvono nel tempo

PROCESSI ALEATORI