Metodologia Statistica Applicata in Ambito Biomedico e Clinico

SECONDA PARTE

Mauro Gambaccini

Anno accademico 2018/19

Probabilità

La probabilità può assumere valori tra 0 1

0 -> certezza che l'evento non si verifichi

1 -> certezza che l'evento si verifichi

COMPLEMENTARITA' DELLA PROBABILITA'

Se p_A è la probabilità che si verifichi l'evento A allora la probabilità che lo stesso evento (evento A) non si verifichi è $(1 – p_A)$

Probabilità che si verifichi A → p_A

Probabilità che non si verifichi A → $(1 – p_A)$

ADDITTIVITA' DELLA PROBABILITA'

Se p_A è la probabilità che si verifichi l'evento A

e p_B è la probabilità che si verifichi l'evento B

allora la probabilità che si verifichi l'evento A o l'evento B è la somma delle rispettive probabilità → $p_{AoB} = p_A + p_B$

Poichè la probabilità è un numero positivo il risultato p_{AoB} sarà certamente maggiore sia di p_A che p_B

ESEMPIO DELLA ADDITTIVITA' DELLA PROBABILITA'

Monteta

Evento [testa] p_t = ½ = 0.5

Evento [croce] p_c = ½ = 0.5

Evento [testa o croce] p_{toc} = 0.5 + 0.5 = 1 → certezza

PROBABILITA' CONDIZIONATA

Se p_A è la probabilità che si verifichi l'evento A
e p_B è la probabilità che si verifichi l'evento B

allora la probabilità che si verifichi l'evento A e l'evento B è data dal prodotto delle rispettive probabilità → $p_{AeB} = p_A \times p_B$

PRODOTTO DELLA PROBABILITA'

Poichè la probabilità è un numero minore di uno il risultato p_{AeB} sarà certamente minore sia di p_A che p_B

ESEMPIO DELLA PRODOTTO DELLA PROBABILITA'

Monteta

Evento [testa] $p_t = ½ \quad = 0.5$

Evento [croce] $p_c = ½ \quad = 0.5$

Evento [testa e testa] in due lanci successivi

$p_{tt} = 0.5 \times 0.5 = 0.25 \rightarrow ¼$ uno ogni 4 prove

PROBABILITA' TESTA E CROCE DI 1 e 2 LANCI DI UNA MONETA

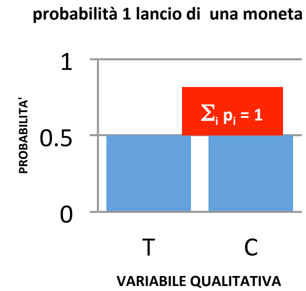
testa croce

p_T T   C p_C

PROBABILITA' 0.5 0.5

1/2 1/2

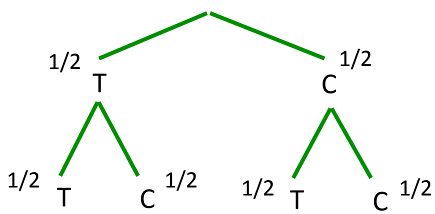
0.5 + 0.5 = 1



lanci

1°

2°



1/2 x 1/2 = 1/4

T T	T C	C T	C C
1/4	1/4	1/4	1/4

T C è equivalente a CT allora le rispettive probabilità si sommano:

T C		C T
1/4	+	1/4
	2/4	

Probabilità Composta

Se

P_A è la probabilità che si verifichi l'evento A

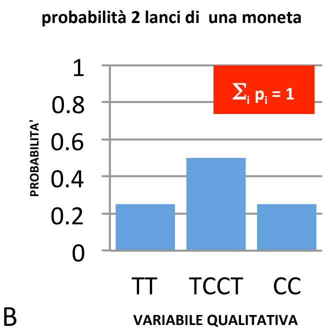
e

P_B è la probabilità che si verifichi l'evento B

allora

$P_A \times P_B$ è la probabilità che si verifichino sia A che B

$P_{AB} = P_A \times P_B$ è la probabilità composta degli eventi A e B



$p_{TT} = p_T \times p_T = 1/2 \times 1/2 = 1/4$ → $p_{TT} = p_T \times p_T = 0.5 \times 0.5 = 0.25$

$p_{CC} = p_C \times p_C = 1/2 \times 1/2 = 1/4$ → $p_{CC} = p_C \times p_C = 0.5 \times 0.5 = 0.25$

$p_{TC} = p_T \times p_C = 1/2 \times 1/2 = 1/4$

$p_{CT} = p_C \times p_T = 1/2 \times 1/2 = 1/4$

→ $p_{TC} + p_{CT} = 1/4 + 1/4 = 0.25 + 0.25 = 0.5$

$\Sigma_i\ p_i = 1$

PROBABILITA’ TESTA E CROCE DI 3 LANCI DI UNA MONETA

testa croce

T C

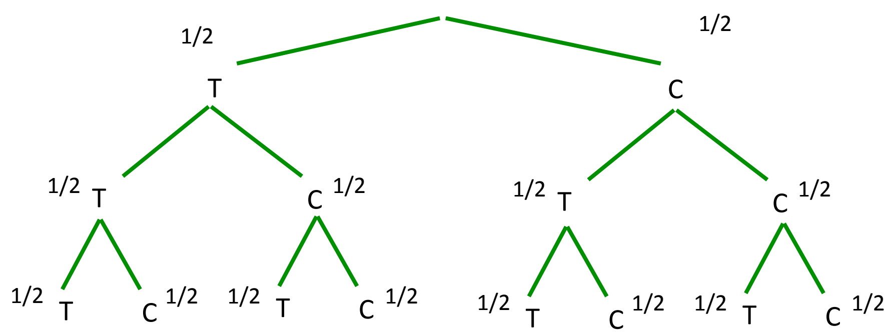
	p_T	p_C
PROBABILITA’	0.5	0.5
	1/2	1/2

lanci

1°

2°

3°

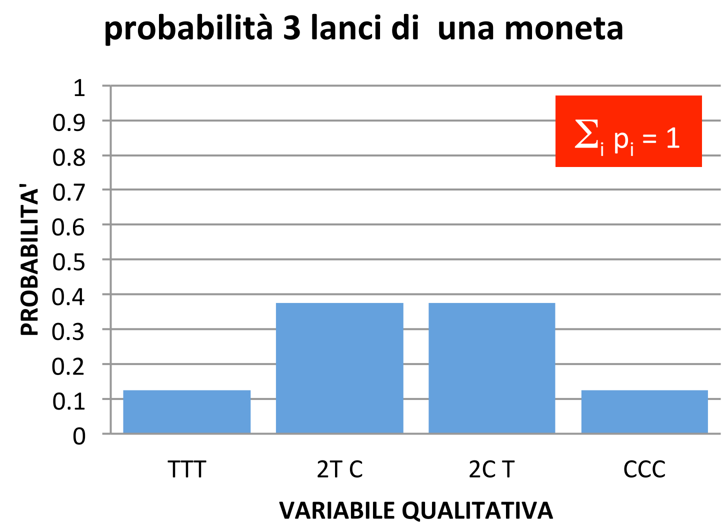


T T T	T TC	TC T	TC C	CTT	CTC	CC T	CC C
1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

1/8 1/8 1/8 1/8 1/8 1/8 1/8 1/8 1/8

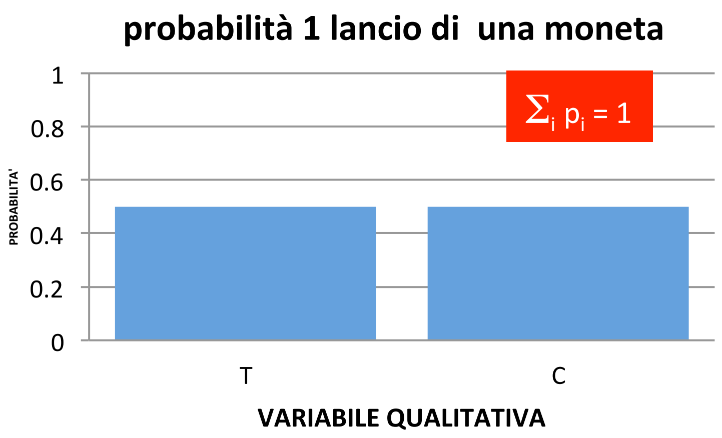
1/8 3/8 3/8 1/8

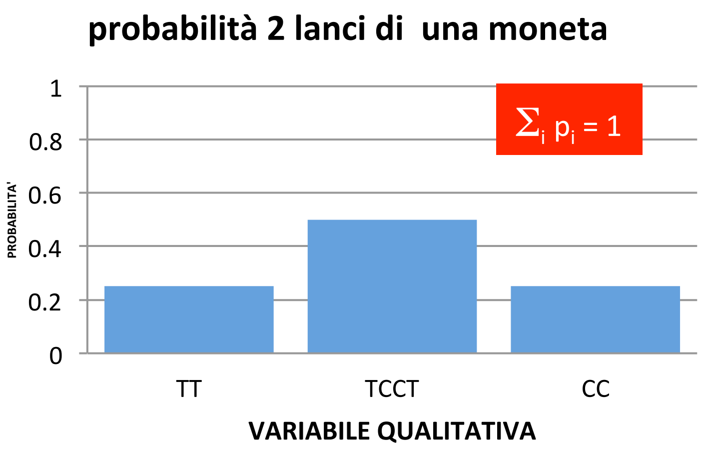
2 TC 2 C T

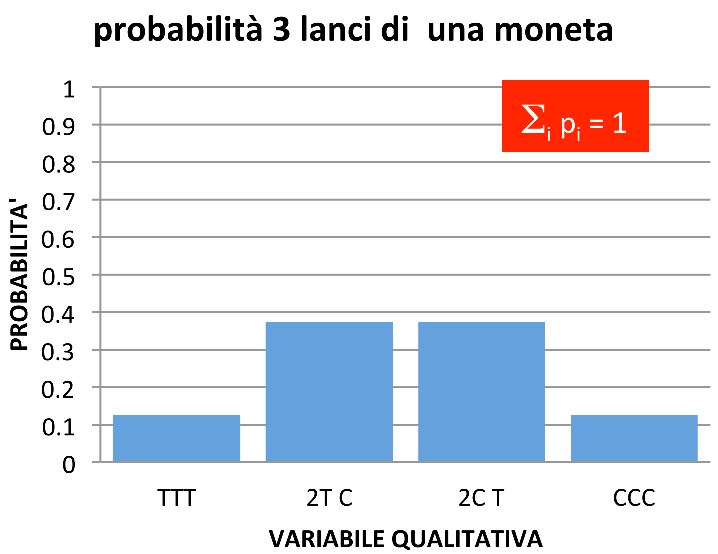


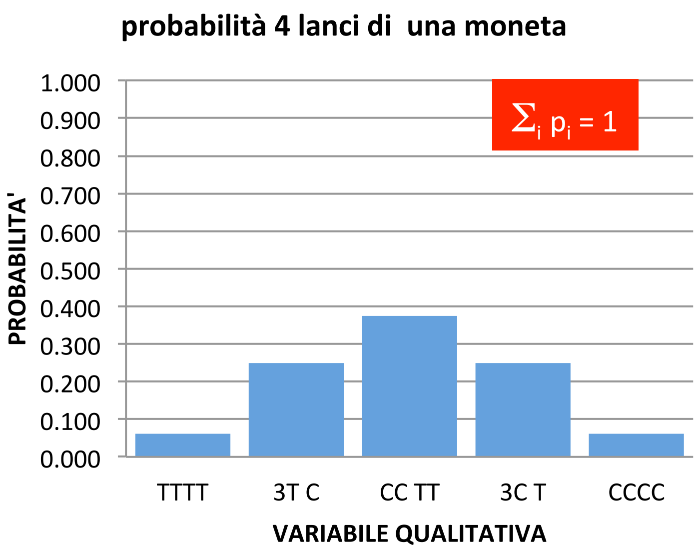
DISTRIBUZIONE DELLE PROBABILITA NEL LACIO DI UNA MONETA

all'aumentare del numero di lanci la distribuzione delle probabilità tende ad assumere una forma a campana









PROBABILITA' NEL LANCIO DI UN DADO

1, 2, 3, 4, 5, 6

ogni numero ha la stessa probabilità di uscire $p_{si} = 1/6 = 0.167$

ogni numero ha anche la stessa probabilità di non uscire $p_{no} = 5/6 = 0.833$

$p_{si} + p_{no} = 1/6 + 5/6 = 6/6 = 1$

$p_{si} + p_{no} = 1$

$p_{no} = 1 - p_{si}$ → La probabilità di non uscire è (1 – la probabilità di uscire)

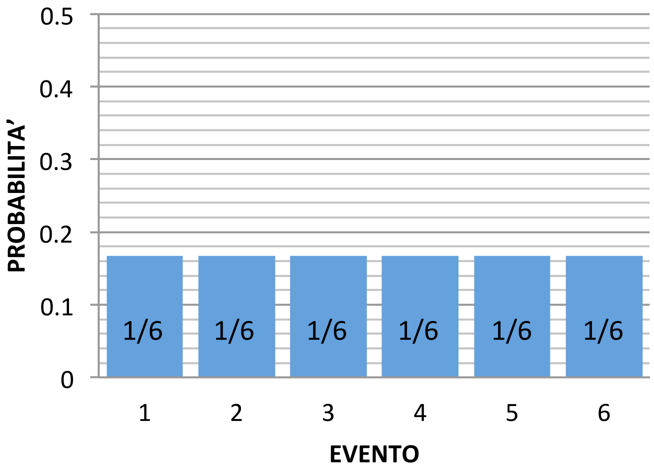
Quindi se la probabilità che un evento si verifichi è → p
La probabilità che quell'evento non si verifichi è → (1 – p)

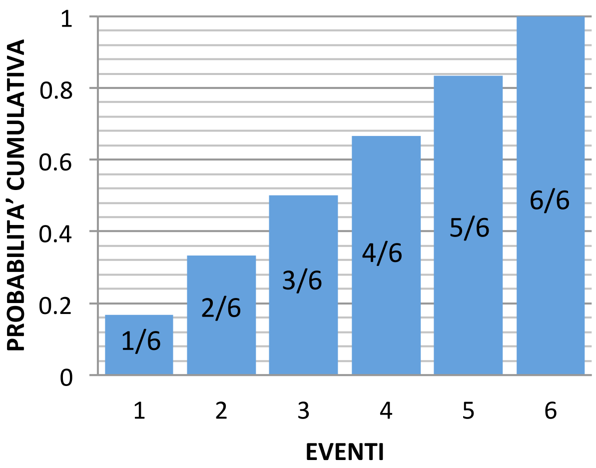
NB: fatto e valido anche per la moneta

PROBABILITA'

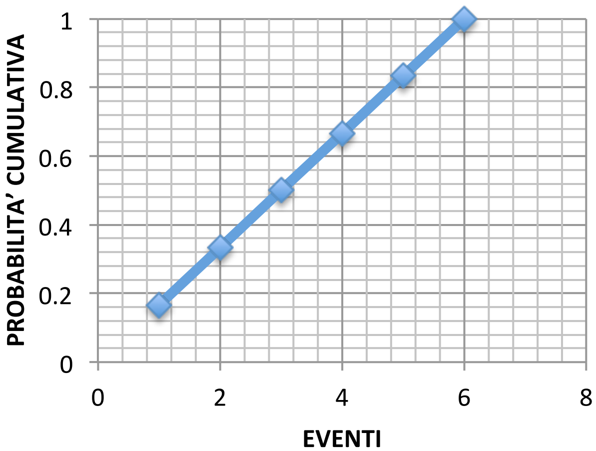
1, 2, 3, 4, 5, 6

PROBABILITA' NEL LANCIO DI UN DADO





Distribuzione della probabilità è piatta
La probabilità cumulativa cresce linearmente

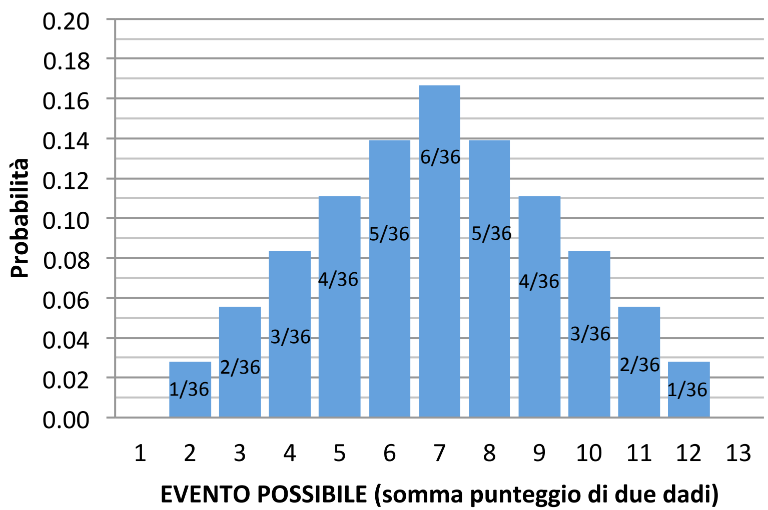


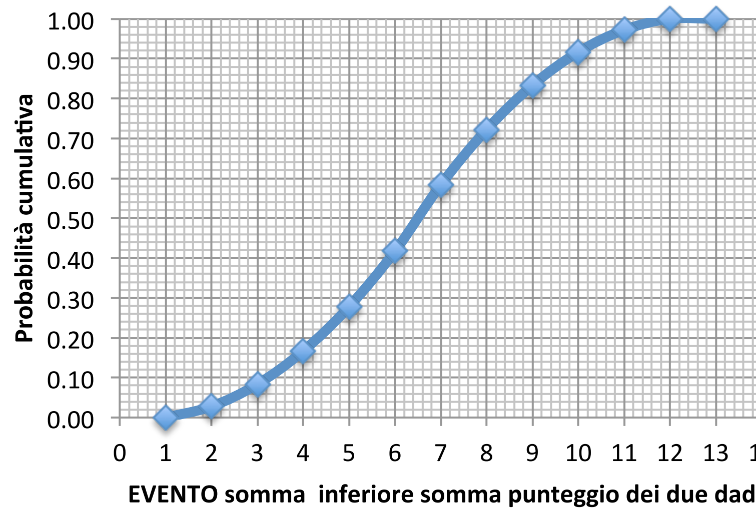
PROBABILITA’ NEL LANCIO DI DUE DADO (somma punteggio)

		DADO 1					
		1	2	3	4	5	6
DADO 2	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

somma	frequenza	Prob.	prob. cum.
1	0	0.00	0.00
2	1	0.03	0.03
3	2	0.06	0.08
4	3	0.08	0.17
5	4	0.11	0.28
6	5	0.14	0.42
7	6	0.17	0.58
8	5	0.14	0.72
9	4	0.11	0.83
10	3	0.08	0.92
11	2	0.06	0.97
12	1	0.03	1.00
13	0	0.00	1.00
totale	36	1	

6 x 6 = 36 dati



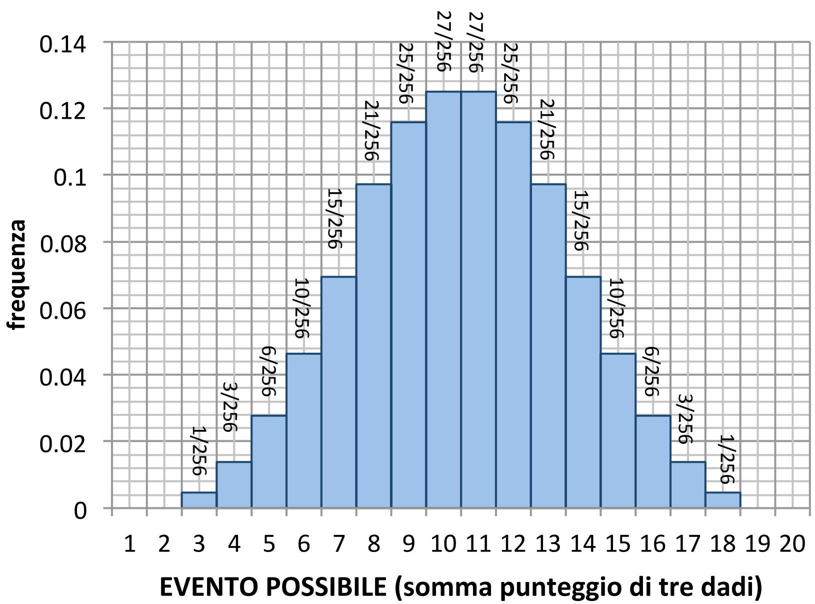


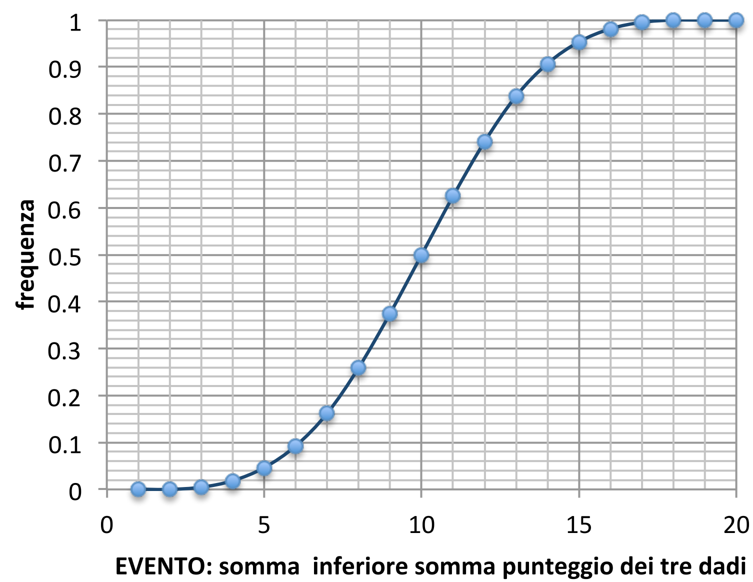
Valore piu probabile 7 p = 0.17 p% = 17%

PROBABILITA’ NEL LANCIO DI TRE DADI
(somma dei punteggi)

		SOMMA DADO 1 E DADO 2 (36 somme)																																			
		2	3	4	5	6	7	3	4	5	6	7	8	4	5	6	7	8	9	5	6	7	8	9	10	6	7	8	9	10	11	7	8	9	10	11	12
DADO 3	1	3	4	5	6	7	8	4	5	6	7	8	9	5	6	7	8	9	10	6	7	8	9	10	11	7	8	9	10	11	12	8	9	10	11	12	13
	2	4	5	6	7	8	9	5	6	7	8	9	10	6	7	8	9	10	11	7	8	9	10	11	12	8	9	10	11	12	13	9	10	11	12	13	14
	3	5	6	7	8	9	10	6	7	8	9	10	11	7	8	9	10	11	12	8	9	10	11	12	13	9	10	11	12	13	14	10	11	12	13	14	15
	4	6	7	8	9	10	11	7	8	9	10	11	12	8	9	10	11	12	13	9	10	11	12	13	14	10	11	12	13	14	15	11	12	13	14	15	16
	5	7	8	9	10	11	12	8	9	10	11	12	13	9	10	11	12	13	14	10	11	12	13	14	15	11	12	13	14	15	16	12	13	14	15	16	17
	6	8	9	10	11	12	13	9	10	11	12	13	14	10	11	12	13	14	15	11	12	13	14	15	16	12	13	14	15	16	17	13	14	15	16	17	18

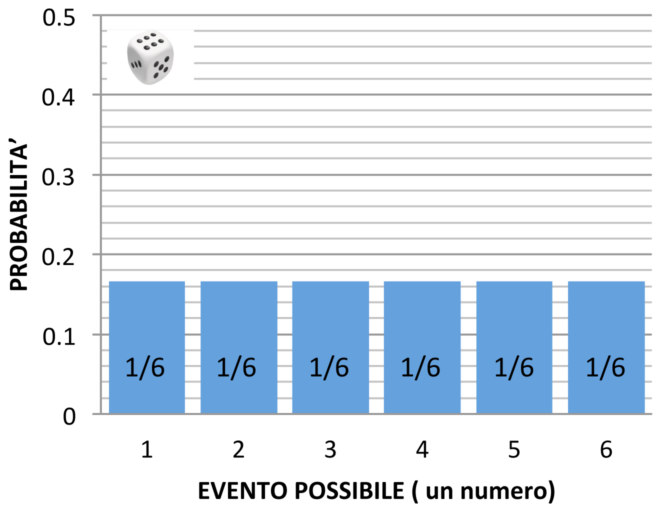
$6 \times 6 \times 6 = 6^3 = 216$ dati

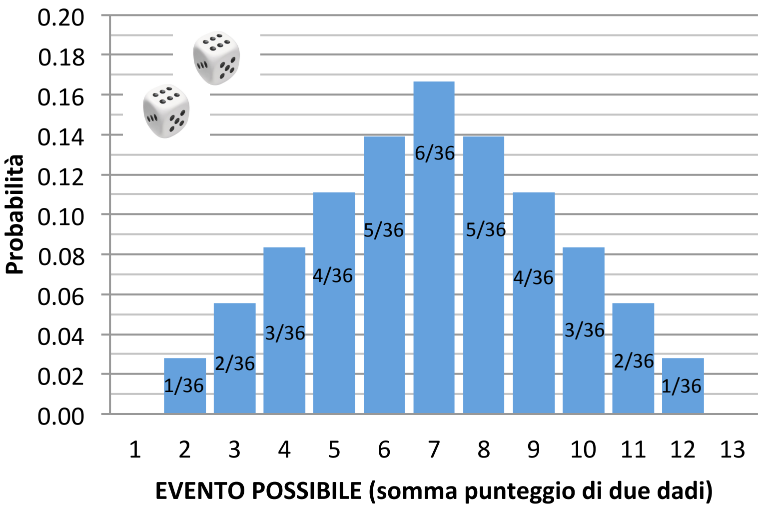


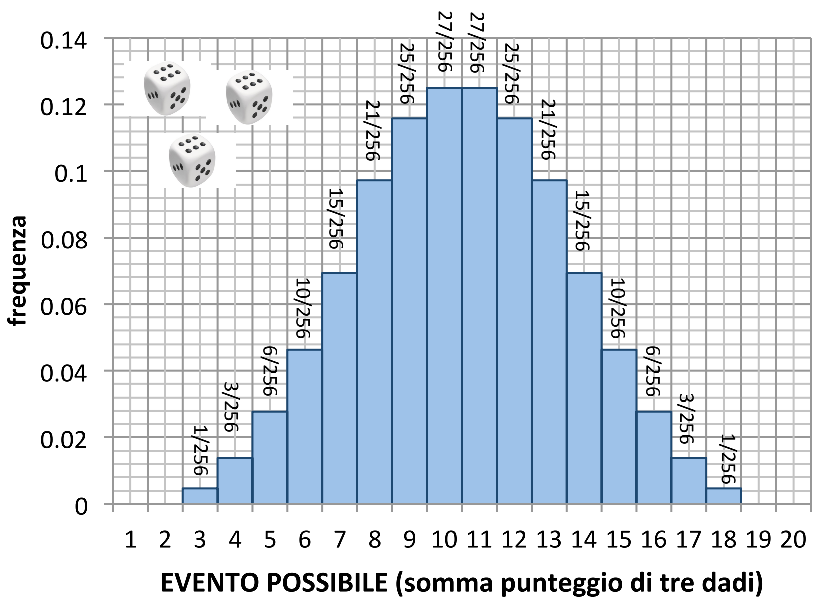


Valori piu probabili **10 e 11** equiprobabili $p = 0.125$ ciascuno, 10 e 11 insieme $p = 0.250$

DISTRIBUZIONE DELLE PROBABILITA NEL LACIO DI DADI

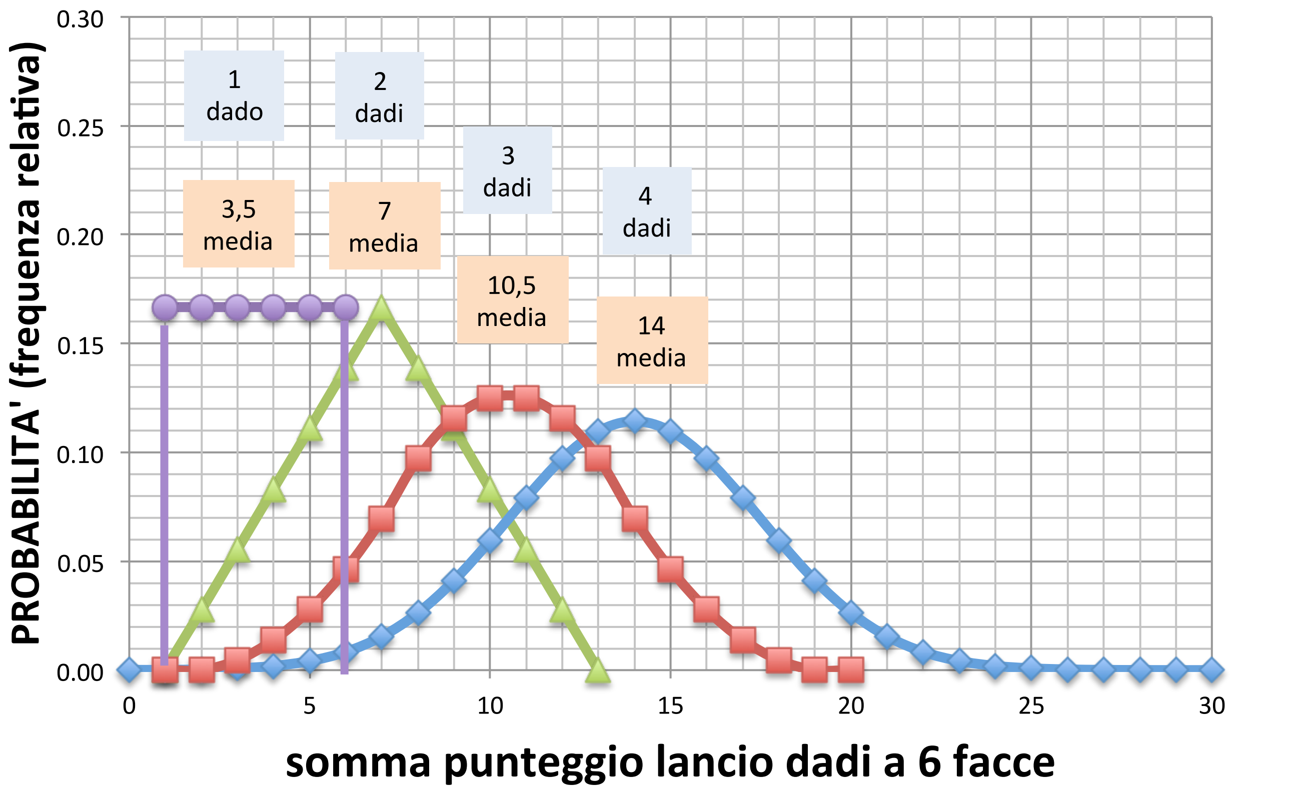






Anche per i dadi all'aumentare del numero di lanci la distribuzione delle probabilità tende ad assumere una forma a campana


distribuzione equiprobabile (probabilità costante): lancio di 1 dado
distribuzione triangolare (probabilità cresce e cala linearmente): lancio di 2 dadi
Distribuzioni a campana (probabilità cresce a cala con un andamento a campana): lancio di 3 o più dadi

All'aumentare del numero di dadi lanciati la distribuzione delle probabilità
Tende alla distribuzione NORMALE ovvero alla finzione di Gauss o Gaussiana

LA DISTRIBUZIONE NORMALE
LA FUNZIONE DI GAUSS
LA GAUSSIANA

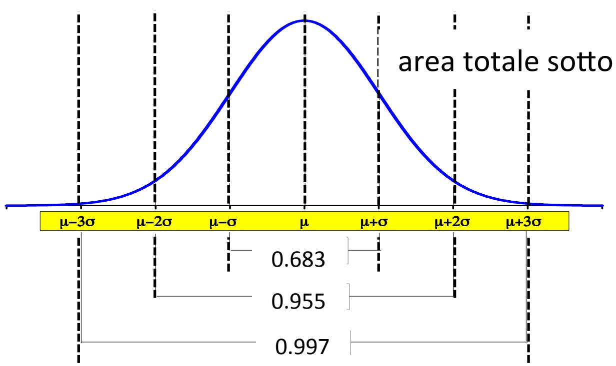
E' la distribuzione a campana più nota e più usata in statistica.
La funzione di Gauss può essere scritta in vari modi due dei quali sono

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

μ = media
σ = d.s. (deviazione standard)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\overline{x}}{\sigma}\right)^2}$$

$\overline{x}$ = media
σ = d.s. (deviazione standard)

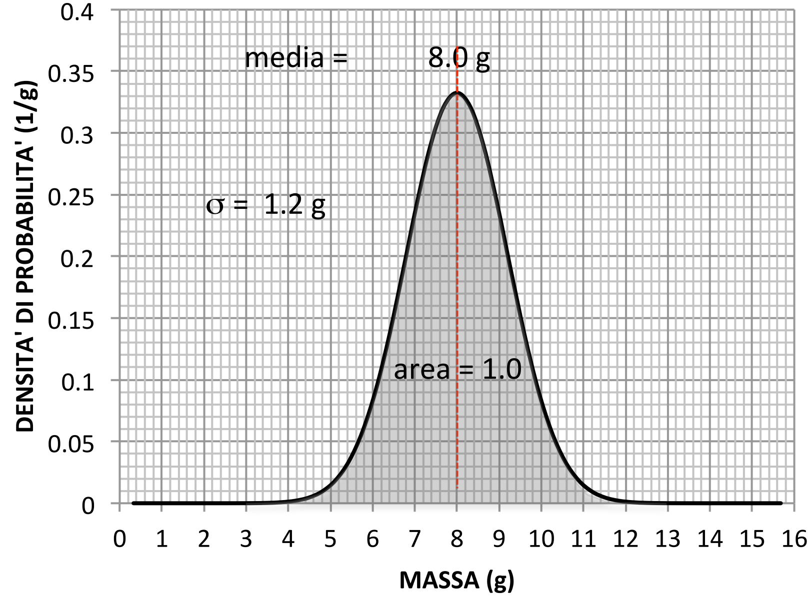


area sotto la curva

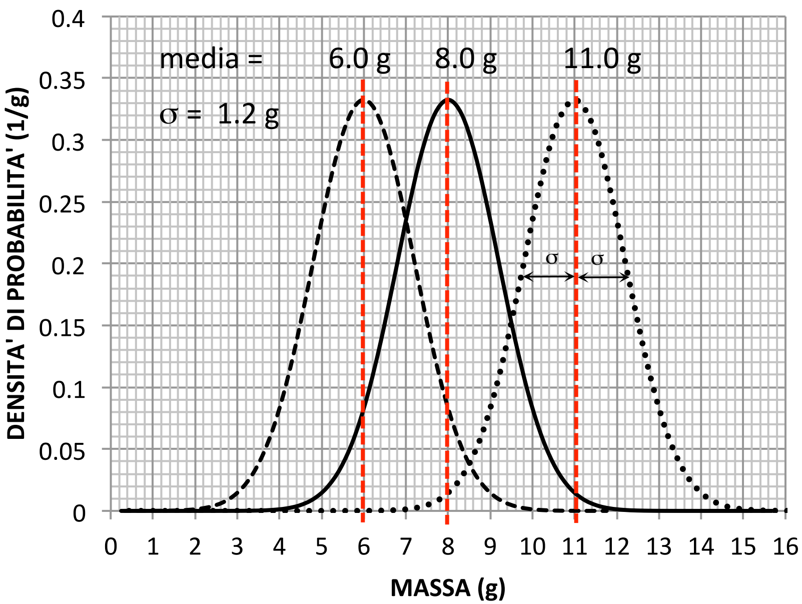
Questa funzione ha alcune caratteristiche molto importanti:

1) Ha il massimo quando la variabile x assume un valore uguale alla media $\bar{x}$;
2) È simmetrica rispetto al massimo
3) È sempre positiva e tende zero per valori di x molto maggiori e molto minori della media.
4) Presenta due flessi a destra e sinistra del massimo nei punti $\bar{x} - \sigma$ e $\bar{x} + \sigma$
5) $\int_{-\infty}^{+\infty} f(x)dx = 1$ ovvero l'area sottesa dalla curva è adimensionale e vale 1.
6) Al variare della media la curva trasla sull'asse delle x.
7) Il valore della deviazione standard σ determina la larghezza della curva

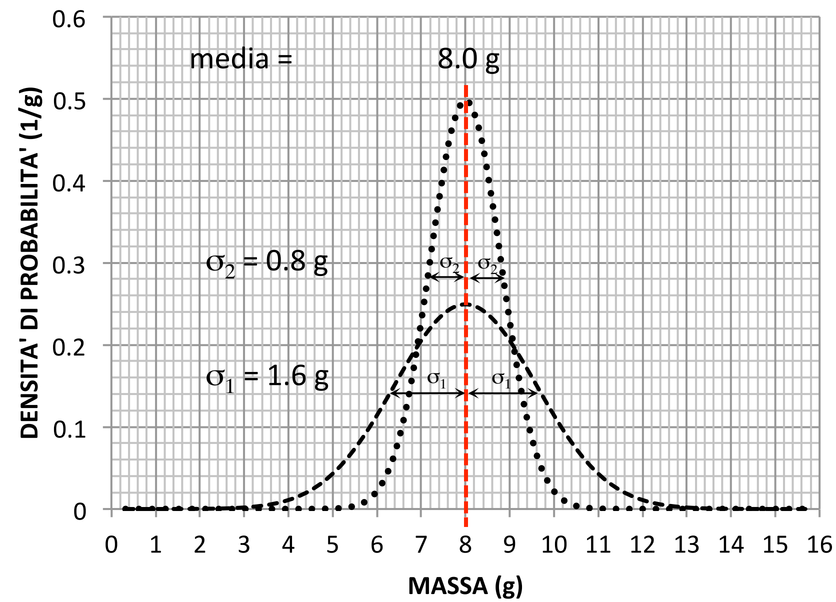
Gaussiana con media di 8 g e d.s. di 1.2 g



Tre Gaussiane con medie diverse e stessa d.s.



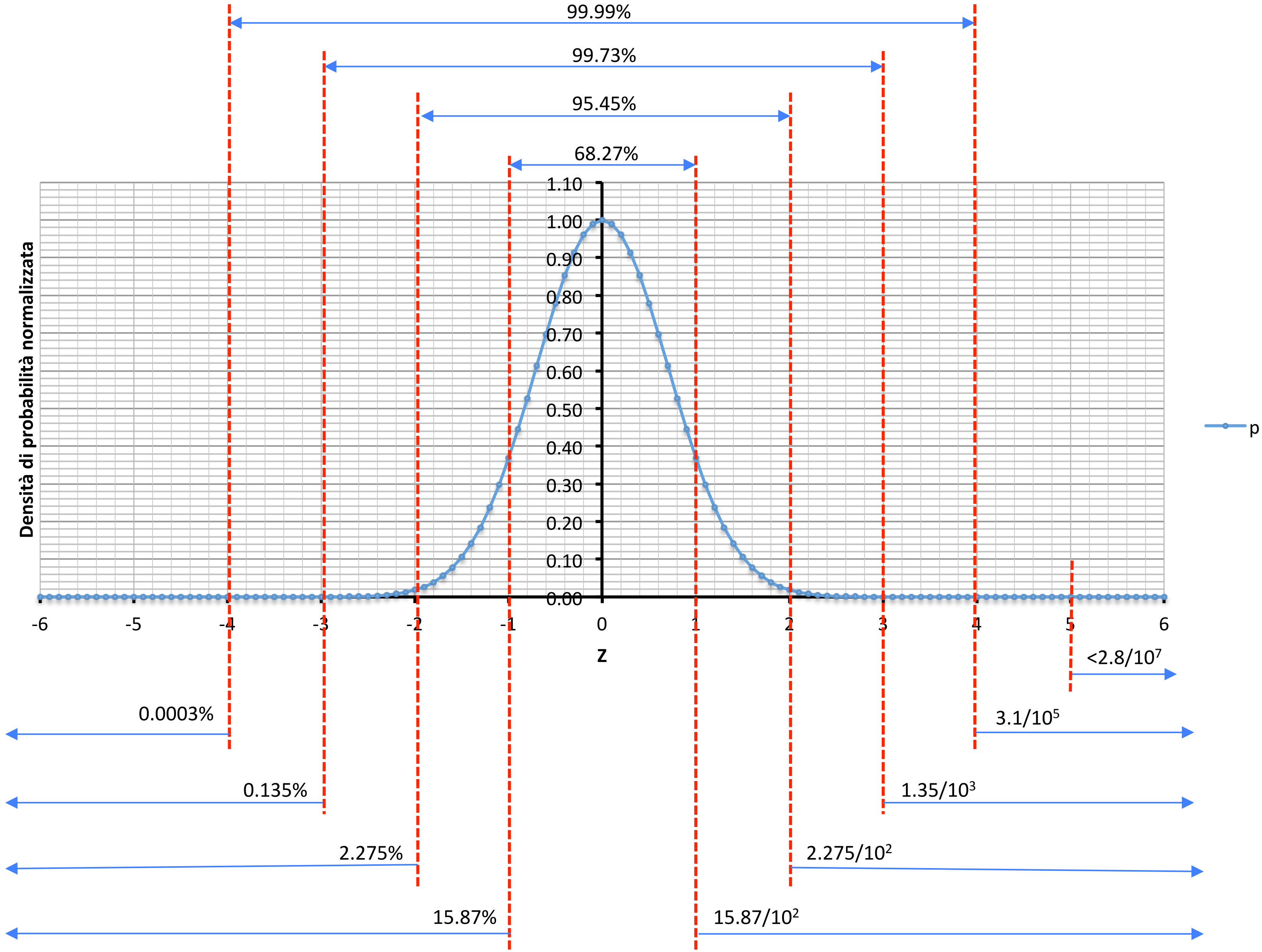
Due Gaussiane con stessa media e diversa d.s.



Gaussiana con media nulla e deviazione standard unitaria

$\mu = 0$

$\sigma = 1$



μ = 4,06 L

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

σ = 0,68 L

μ = media
σ = d.s. (deviazione standard)

Supponiamo di conoscere la media e la deviazione standard di un campione di misure di volume respiratorio μ = 4,06 L σ = 0,68 L
Se inseriamo i rispettivi valori nella formula cosa otteniamo?

La variabile x la media μ e la d.s. σ sono tutte espresse il L (litri) per cui l'esponente della e, essendo il rapporto (L^2/L^2) è adimensionale mentre il coefficiente $\frac{1}{\sigma\sqrt{2\pi}}$ risulta essere espresso il unità (1/L).

Questo sta ad indicare che la funzione di Gauss $f(x)$ essendo espressa in unità (1/L) non è una probabilità ma bensì una **densità di probabilità** e per trasformarla in probabilità sarà necessario moltiplicarla per l' ampiezza di classe Δx che in questo caso è espressa il litri (L)
La probabilità di avere un valore respiratorio compreso tra il valore x ed il valore $x + \Delta x$ è data da: $f(x) \,.\, \Delta x$
il risultato di questo prodotto è adimensionale -> (1/L)x(L)
proprio come deve essere la probabilità ovvero un numero puro (adimensionale)

DALLA DENSITA' DI PROBABILITA'   ALLA PROBABILITA'

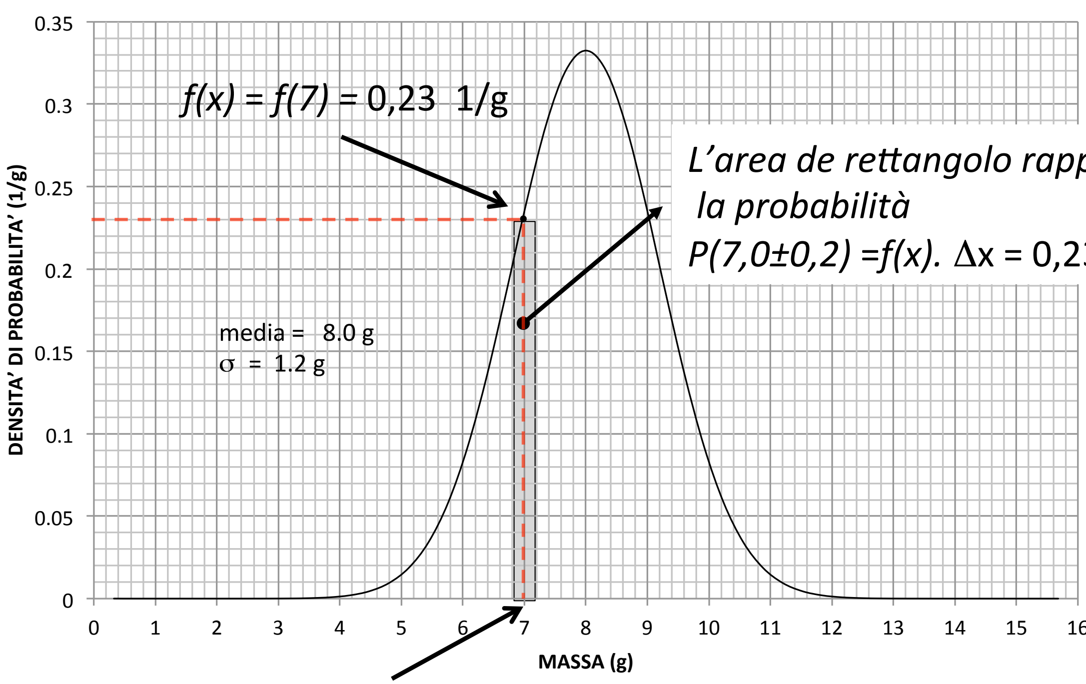
Esempio ci calcolo della probabilità partendo dalla Gaussiana

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

μ = 8.0 g

σ = 1.2 g

x = 7.0 g



L'area de rettangolo rappresenta la probabilità P(7,0±0,2) =f(x). Δx = 0,23 . 0,4 = 0,092

x = 7 g   Δx = 0,4 g

DALLA PROBABILITA' ALLA FREQUENZA

Abbiamo ottenuto una probabilità di 0,092 ovvero 9,2% .

Questo numero è la probabilità che un seme scelto a caso in un campione "ipotizzato gaussiano con media 8.0 g e σ 1.2 g" abbia una massa compresa tra 6,8 g e 7,2 g (estremi della classe pari all' ampiezza utilizzata nel calcolo della probabilità: 0,4 g)

Supponiamo ora che il campione sia costituito da 1000 semi,

allora,

tra quegli estremi (6,8 g e 7,2 g) dovremmo trovare 1000 x 0.092 = 92 semi

92 è la frequenza della classe compresa tra 6,8 g e 7,2 g per il campione ipotizzato gaussiano

Il prodotto del numero totale di elementi del campione N e della probabilità della classe i-esima p_i , fornisce la frequenza della classe i-esima:

$$f_i = N \times p_i$$

*Conoscendo media μ , deviazione standard σ e numero di elementi **N** di un campione è possibile ricostruire la distribuzione delle frequenze di quel campione gaussiano avente la stessa μ , σ ed **N** per una ampiezza di classe scelta a piacere (qualsiasi)*

DALLA DENSITA' DI PROBABILITA' ALLA PROBABILITA'

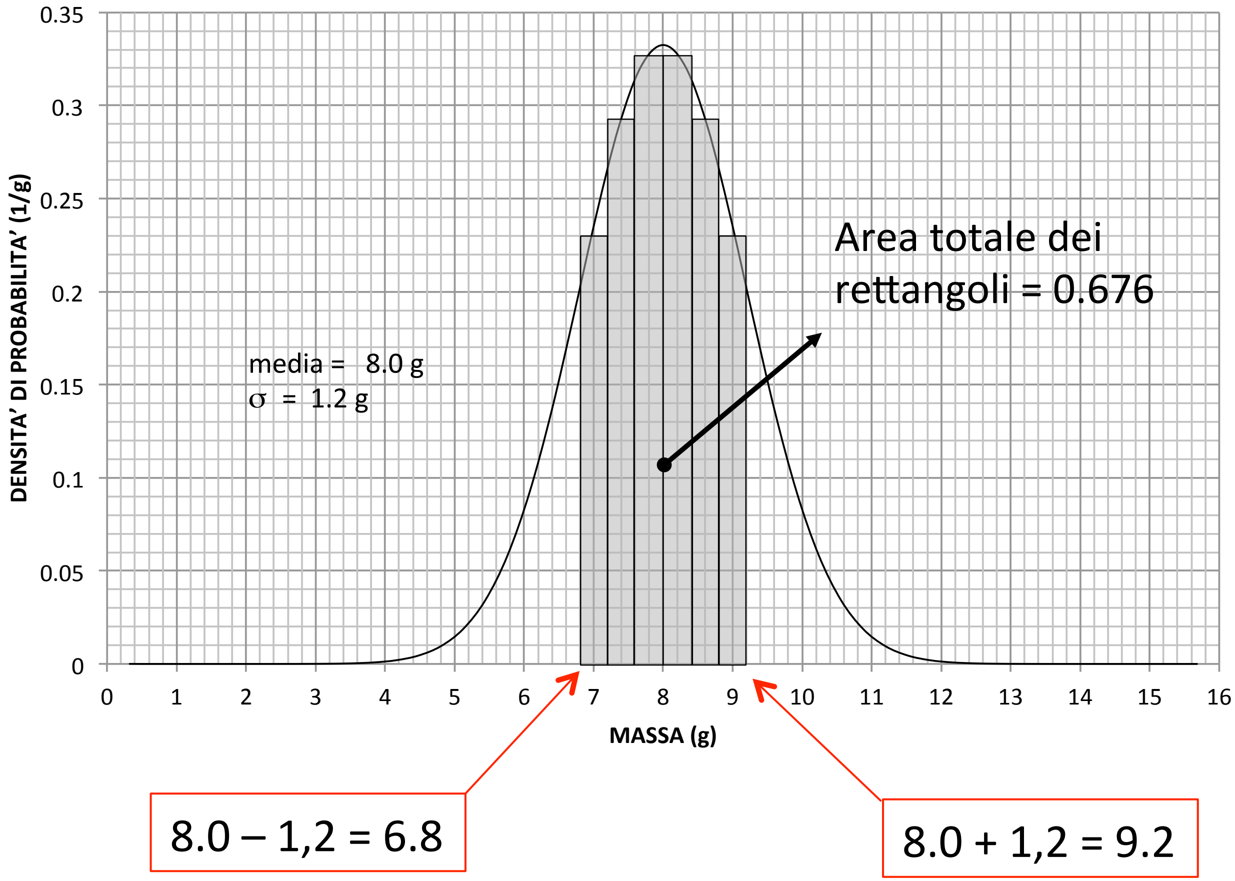
Esempio con calcolo approssimato dell'area di una parte di Gaussiana

Esempio di calcolo approssimato delle frequenze che i dati del campione assumano valori compresi tra 8.0 g ± 1.2 g

Il valore viene ottenuto sommando tutte le aree dei rettangoli evidenziati in figura.

Ogni rettangolo ha una larghezza di base pari a 0,4 g

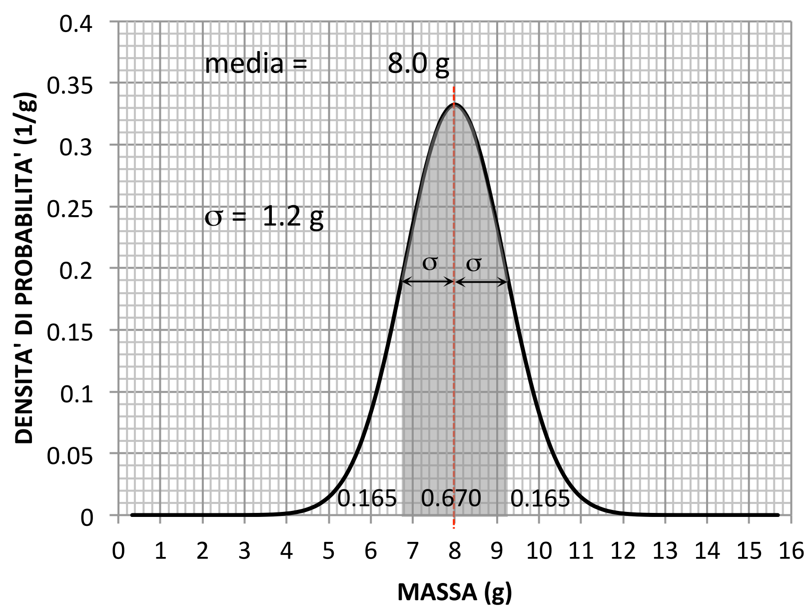
Frequenza o probabilità = (0.23 + 0.29 + 0.325 + 0.325 + 0.29 + 0.23)• 0,4 = 0.676



DALLA DENSITA' DI PROBABILITA' ALLA PROBABILITA'

Esempio ci calcolo esatto dell'area di una parte di Gaussiana

$$\int_{-\infty}^{+6.8} f(x)dx = 0.165 \quad ; \quad \int_{+6.8}^{+9.2} f(x)dx = 0{,}670 \quad ; \quad \int_{+9.2}^{+\infty} f(x)dx = 0.165$$



Risultati ottenuti con il calcolo approssimato	0.162	0.676	0.162
	≈ - 2%	≈ + 1%	≈ - 2%

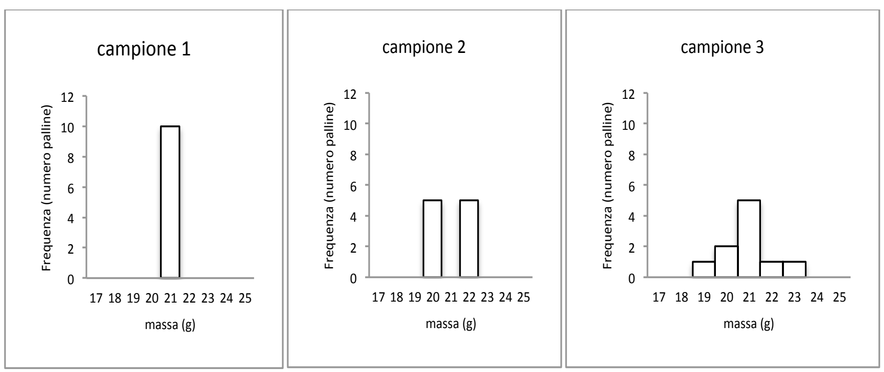
TEST PER VERIFICARE SE UN CAMPIONE DI DATI PROVIENE DA UNA POPLAZIONE NORMALE (Gaussiana)

Supponiamo di avere tre campioni ognuno di 10 palline aventi massa rispettivamente:

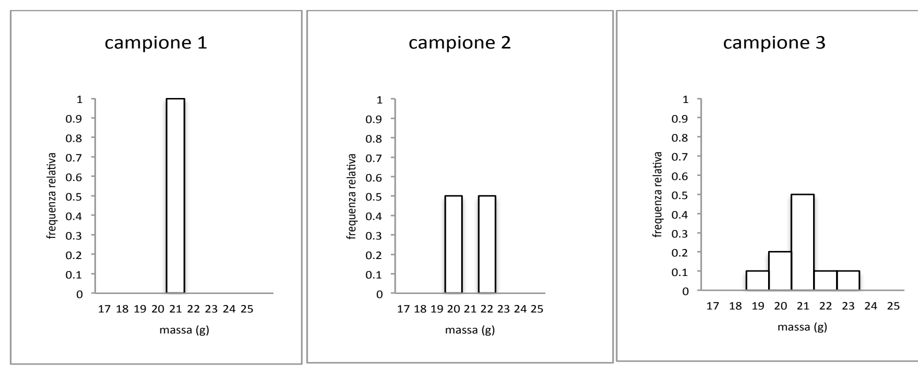
Campione 1: 10 palline con lo a stessa massa pari a 21 g
Campione 2: 5 palline con massa 18 g e 5 palline con massa 24 g
Campione 3: 1 pallina 18 g, 2 palline 19 g, 3 palline 21 g, una 22 g, una 23 g e due 24 g
In numeri:
Campione 1 : 21 g ; 21 g ; 21 g ; 21 g ; 21 g ; 21 g ; 21 g ; 21 g ; 21 g ; 21 g ;
Campione 2 : 18 g ; 18 g ; 18 g ; 18 g ; 18 g ; 24 g ; 24 g ; 24 g ; 24 g ; 24 g ;
Campione 3 : 18 g ; 19 g ; 19 g ; 21 g ; 21 g ; 21 g ; 22 g ; 23 g ; 24 g ; 24 g ;

media Campione 1 : 21 g   **ds** Campione 1 : 0.0 g
media Campione 2 : 21 g   **ds** Campione 2 : 1.1 g
media Campione 3 : 21 g   **ds** Campione 3 : 1.1 g

distribuzione delle frequenze dei tre campioni



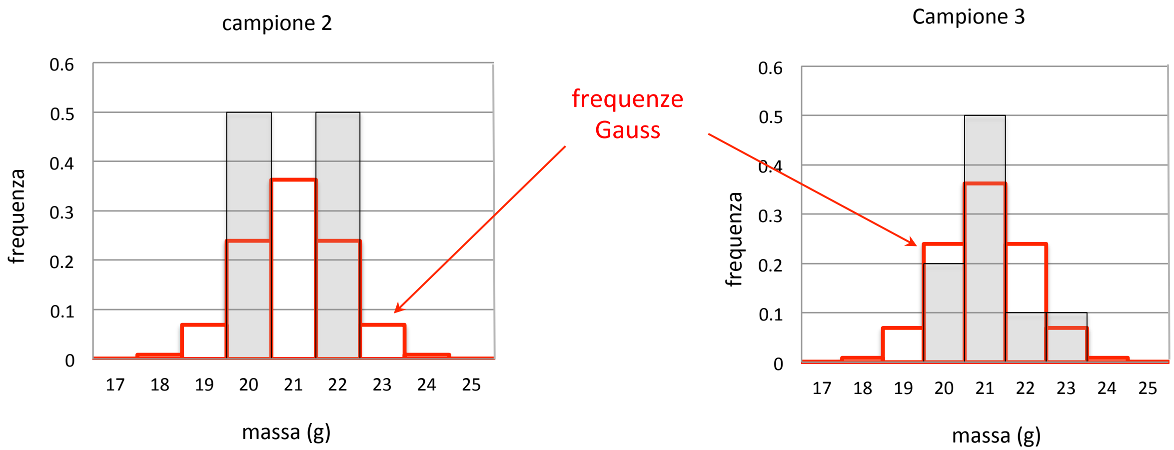
distribuzione delle frequenze relative dei tre campioni



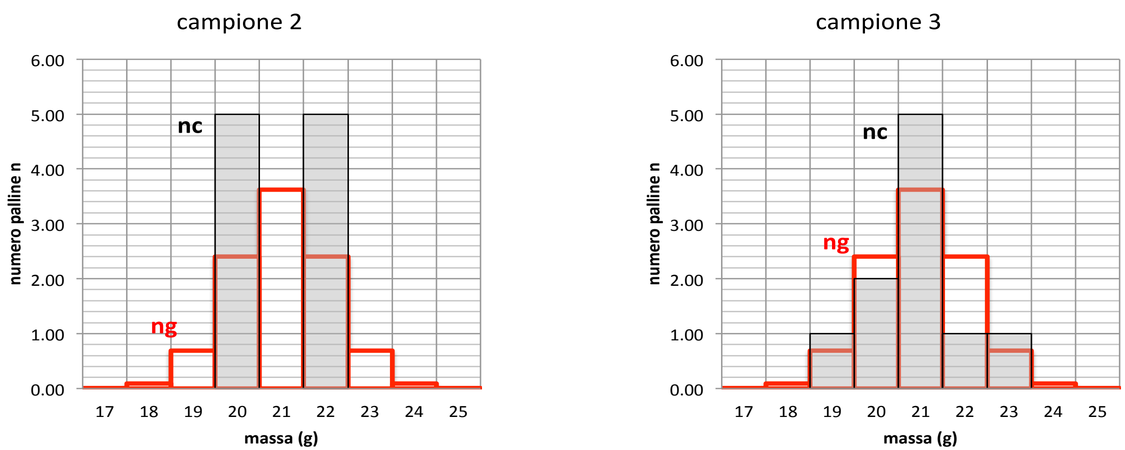
Quale di questi tre campioni ha la probabilità maggiore di provenire da una popolazione di palline la cui massa abbia una distribuzione normale?
Probabilmente la 3, come lo possiamo dimostrare?

CONFRONTO TRA LE DISTRIBUZIONI DELLE FREQUENZE RELATIVE DEI CAMPIONI 2 e 3 CON QUELLE FORNITE DALLA FUNZIONE DI GAUSS CON MEDIA 21 g e d.s. 1.1 g

distribuzione delle frequenze relative dei campioni 2 e 3



distribuzione delle frequenze relative dei campioni 2 e 3

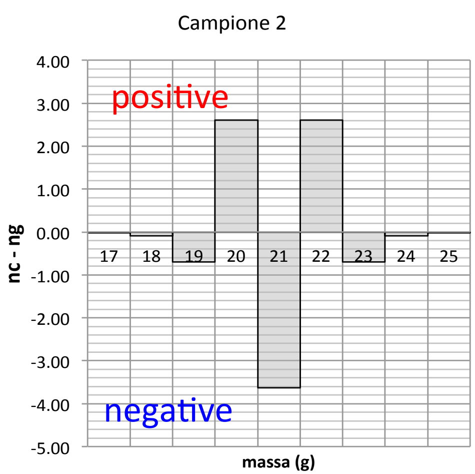


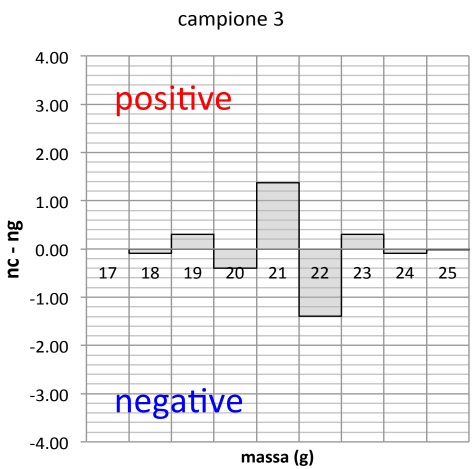
Per passare dalle frequenze relative allle frequenze ovvero al numero di palline atteso in ogni classe è sufficiente moltiplicare la frequenza relativa per il numero totale di palline del campione (10 nel nostro caso)

Il campione 3 si sovrappone di più del campione 2
COME POSSIAMO QUANTIFICARE QUESTE DIFFERENZE?

FACCIAMO LA DIFFERENZA TRA LE FREQUENZE DEL CAMPIONE E QUELLE PREVISTE DALLA DISTRIBUZIONE DI Gauss

differenza del numero di campioni osservati (**nc**) e quelle previsti dalla distribuzione di Gauss (**ng**)





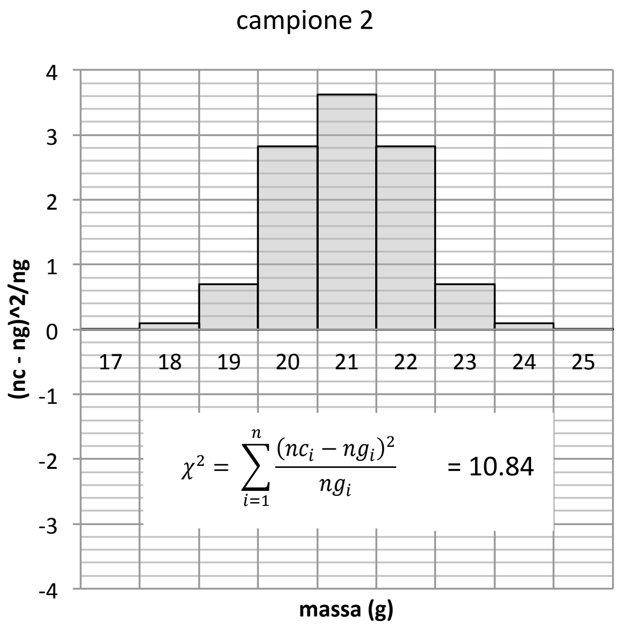
Sommare le differenze ci porterebbe a valori nulli come nel caso del calcolo della deviazione standard per cui dobbiamo elevare la quadrato le differenze introducendo il parametro

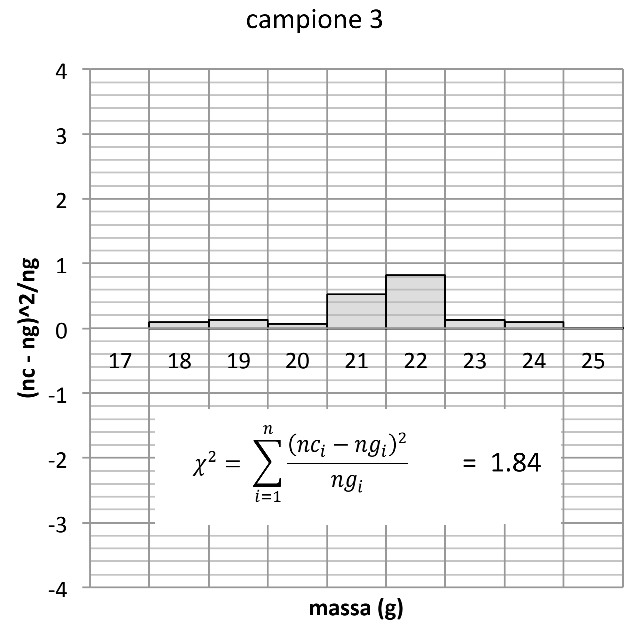
$$C_i = \frac{(nc_i - ng_i)^2}{ng_i}$$

Quadrato della differenza del numero di campioni osservati meno quelli previsti dalla distribuzione normale diviso per numero di campioni previsti dalla distribuzione di Gauss

SOMMANDO I VALORI c_i DI TUTTE LE CLASSI OTTENIAMO IL PARAMETRO *“ CHI QUADRO”* χ^2

$$C_i = \frac{(nc_i - ng_i)^2}{ng_i} \qquad\qquad \chi^2 = \sum_{i=1}^{n} \frac{(nc_i - ng_i)^2}{ng_i}$$





Quando il valore del χ^2 è minore o uguale al numero delle classi (nel nostro caso 7) l’ipotesi che il campione provenga da una popolazione di palline la cui massa abbia una distribuzione normale può essere accettata. Se invece è maggiore l’ipotesi viene scartata.