Distribuzioni teoriche di probabilità: distribuzione binomiale e distribuzione normale

- Prof. Giuseppe Verlato
- Sezione di Epidemiologia e Statistica Medica, Università di Verona

media e varianza

campione di numerosità n — variabile casuale n?

media $\bar{x} = \Sigma x / n$

$\mu = E(X) = \Sigma x * p$ **var.discreta**

$\mu = E(X) = \int x * f(x)dx$ **var. continua**

varianza $s^2 = \Sigma(x-\bar{x})^2/(n-1)$

$s^2 = \dfrac{\Sigma x^2 - (\Sigma x)^2/n}{(n-1)}$

$s^2 = E(X-\mu)^2$

$s^2 = E(X^2) - \mu^2$

DISTRIBUZIONI TEORICHE DI PROBABILITA'

DISTRIBUZIONE		attesa (media) $E(X)$	varianza $E[X-E(X)]^2$
binomiale	$p(x)=\dfrac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$	$n\pi$	$n\pi(1-\pi)$
di Poisson	$p(x)=\dfrac{\mu^x e^{-\mu}}{x!}$	$\mu=n\pi$	$\mu=n\pi$
normale	$f(x)=\dfrac{1}{\sigma\sqrt{(2\pi)}}e^{-(x-\mu)^2/2\sigma^2}$	μ	σ^2

Esercizio sulla distribuzione binomiale

Variabile casuale = numero di aborti spontanei su 4 gravidanze.

n aborti	0	1	2	3	4
n donne	24	28	7	5	6

Gli aborti spontanei sono distribuiti “a caso” fra le varie donne? o tendono a concentrarsi in alcune di esse?

↓

Se gli aborti si distribuiscono a caso fra le varie donne, la variabile casuale “numero di aborti spontanei su 4 gravidanze” deve seguire la distribuzione binomiale.

Esercizio sulla distribuzione binomiale

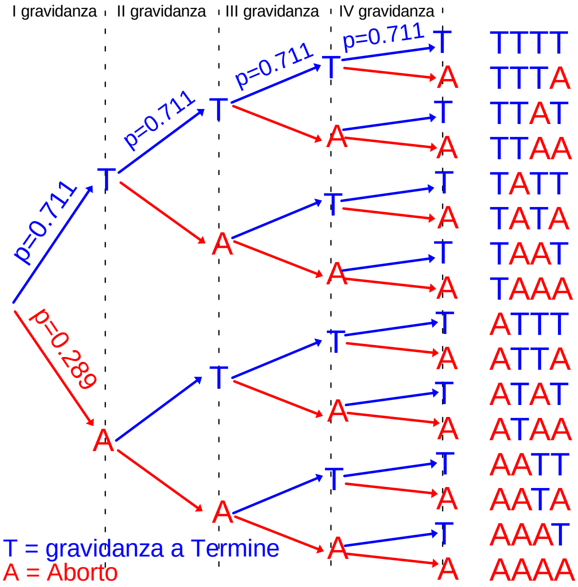
Assunzioni alla base della distribuzione binomiale	Nell'esempio:
La variabile deve esprimere il numero di successi (o insuccessi) su n prove	numero di aborti su 4 gravidanze
Ogni singola prova deve ammettere solo 2 possibili risultati, successo (0) e insuccesso (1)	ogni gravidanza può esitare in un aborto spontaneo o in un neonato
La probabilità deve rimanere costante in tutto il campione	La probabilità di aborto deve rimanere costante fra tutte le donne
Il risultato di una prova non deve essere influenzato dal risultato delle prove precedenti	La probabilità di aborto deve essere indipendente dall'esito delle gravidanze precedenti

Esercizio sulla distribuzione binomiale

		TOTALE
donne =	24 + 28 + 7 + 5 + 6	= 70
gravidanze =	70 * 4	= 280
N aborti =	24*0 + 28*1 + 7*2 + 5*3 + 6*4	= 81

p(aborto) = 81/280 = 0.289

Variabile = n. di aborti su 4 gravidanze

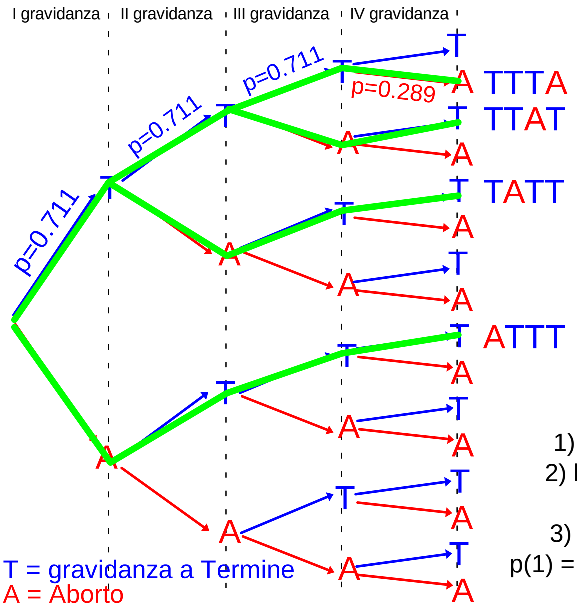


Qual e' la probabilita'
di avere ZERO aborti,
p(0), p(TTTT)?

Se gli eventi fossero
EQUIPROBABILI,
p(0) = 0.5*0.5*0.5*0.5
= 0.0625 = 1/16

Dal momento che
1) p(T)=0.711,
2) le prove (gravidanze)
sono fra loro indipendenti,
p(TTTT) =
=0.711*0.711*0.711*0.711=
= 0.255

Variabile = n. di aborti su 4 gravidanze

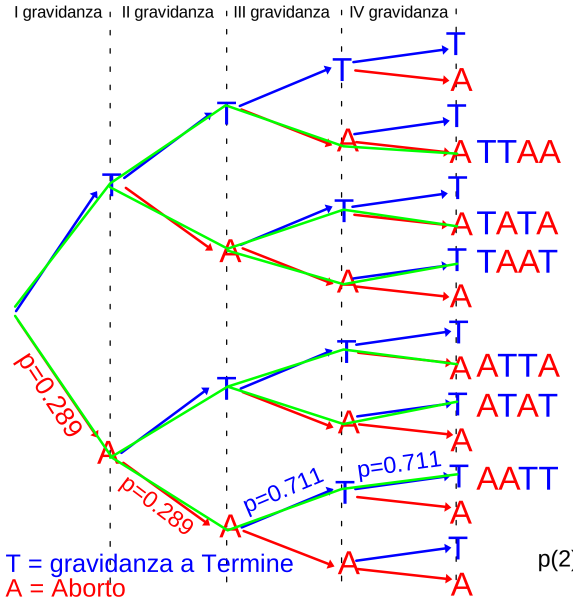


Qual e' la probabilita'
di avere UN aborto,
p(1)?

Se gli eventi fossero
EQUIPROBABILI,
p(1) = 4/16 = 1/4

Dal momento che
1) p(A)=0.289, p(T)=0.711
2) le prove (gravidanze) sono
fra loro indipendenti,
3) i percorsi possibili sono 4,
p(1) = 4*0.289*0.711*0.711*0.711 =
= 0.415

Variabile = n. di aborti su 4 gravidanze



Qual e' la probabilita' di avere DUE aborti, p(2)?

Se gli eventi fossero EQUIPROBABILI, p(2) = 6/16 = 3/8

Dal momento che
1) p(A)=0.289, p(T)=0.711
2) le prove (gravidanze) sono fra loro indipendenti,
3) i percorsi possibili sono 6,
p(2) = 6*0.289*0.289*0.711*0.711 =
= 0.253

Variabile = n. di aborti su 4 gravidanze

p(0) = (0.711*0.711*0.711*0.711) = 0.255

p(1) = 4 * 0.289 * (0.711*0.711*0.711) = 0.415

p(2) = 6 * (0.289*0.289) * (0.711*0.711) = 0.253

Esiste una formula unificante?

$$p(x) = \frac{n!}{x!(n-x)!} \, p^{x} \, (1-p)^{n-x}$$

numero di percorsi possibili

p(aborto)

p(grav. a termine)

N.B. 4! = 4 fattoriale = 4* 3 * 2 * 1 = 24
0! = 1

In una distribuzione binomiale $p(x)=\frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$

In questo caso n=4 p=0,289

$p(0)=\frac{4!}{0!\,4!}\,0{,}289^0(1-0{,}289)^4 = 0{,}255$

$p(1)=\frac{4!}{1!\,3!}\,0{,}289^1(1-0{,}289)^3 = 0{,}415$

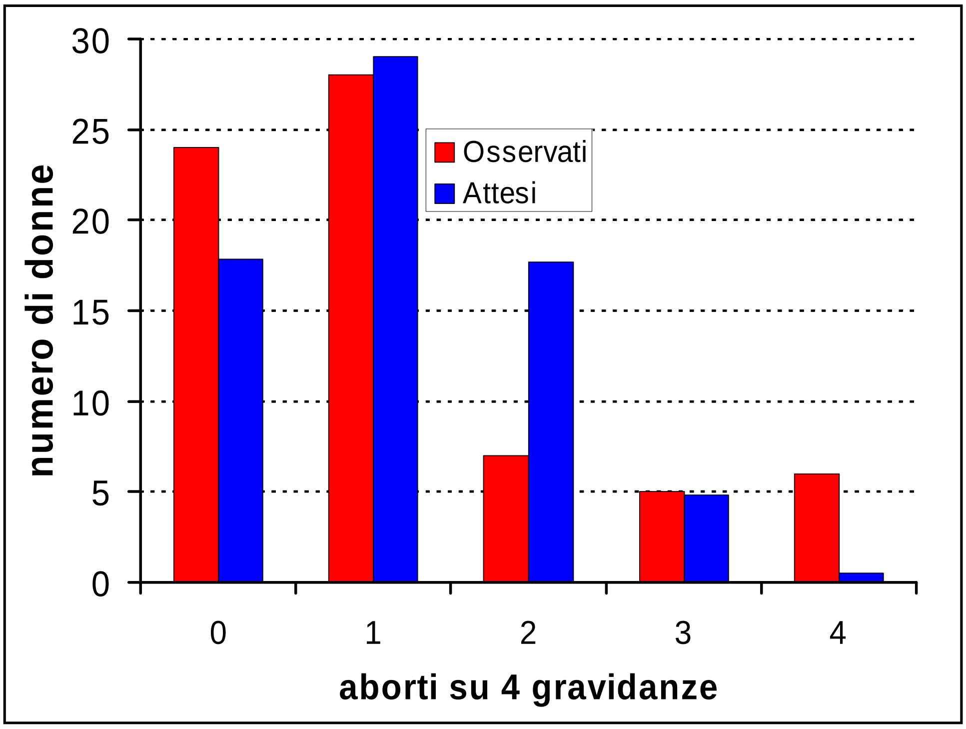
$p(2)=\frac{4!}{2!\,2!}\,0{,}289^2(1-0{,}289)^2 = 0{,}253$

$p(3)=\frac{4!}{3!\,1!}\,0{,}289^3(1-0{,}289)^1 = 0{,}069$

$p(4)=\frac{4!}{4!\,0!}\,0{,}289^4(1-0{,}289)^0 = 0{,}007$

numero di aborti su 4 gravidanze	osservati	attesi se gli aborti fossero distribuiti a caso
0	24	0,255*70 = 17,85
1	28	0,415 * 70 = 29,05
2	7	0,253 * 70 = 17,71
3	5	0,069 * 70 = 4,83
4	6	0,007 * 70 = 0,49

I valori osservati si discostano dai valori attesi: gli aborti non sono distribuiti a caso fra le varie donne ma tendono a concentrarsi in alcune di esse, ad esempio per ipoplasia uterina o per collo dell’utero beante.



Test del chi-quadrato

$$\chi^2 = \Sigma\,(\text{osservati} - \text{attesi})^2 / \text{attesi}$$

H_0: i dati osservati seguono la distribuzione binomiale
H_1: i dati osservati non seguono la distribuzione binomiale

Livello di significatività = 5%
Gradi di libertà =n° celle - n° parametri = 5-2 = 3

Soglia critica = $\chi^2_{3,\,0,05}$ = 7,81

$$\chi^2 = \frac{(24-17{,}9)^2}{17{,}9} + \frac{(28-29{,}1)^2}{29{,}1} + \frac{(7-17{,}7)^2}{17{,}7} + \frac{(5-4{,}8)^2}{4{,}8} + \frac{(6-0{,}5)^2}{0{,}5}$$

$$= 2{,}12 + 0{,}04 + 6{,}48 + 0{,}01 + 61{,}96 = 70{,}60$$

χ^2 osservato > soglia critica
70,60 7,81

Rifiuto H_0

dalla distribuzione BINOMIALE alla distribuzione di POISSON

DISTRIBUZIONE BINOMIALE

$$p(x)= \binom{n}{x} \pi^{x} (1-\pi)^{n-x}$$

$$p(x)= \frac{n*(n-1)*...*(n-x+1)}{x!} \left(\frac{\mu}{n}\right)^{x} \left(1-\frac{\mu}{n}\right)^{n-x}$$

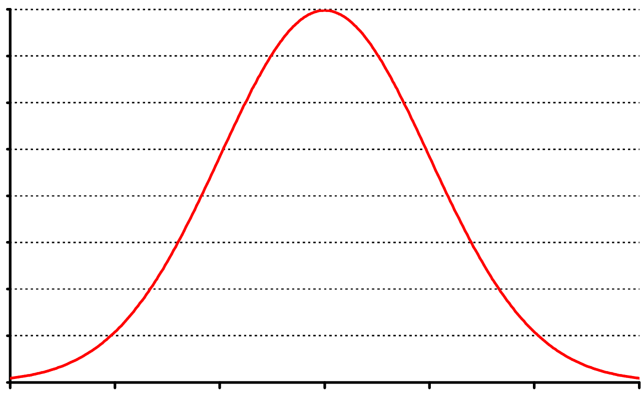
se $n \rightarrow \infty$ e se $x \rightarrow 0$

$$p(x)= \frac{n^{x}}{x!} \frac{\mu^{x}}{n^{x}} \left(1-\frac{\mu}{n}\right)^{n}$$

DISTRIBUZIONE di POISSON

$$p(x)= \frac{\mu^{x}}{x!} e^{-\mu}$$

La distribuzione normale (distribuzione gaussiana, distribuzione degli errori accidentali) occupa un ruolo centrale nell'ambito della statistica medica.



Infatti:

1) Se si sommano 30 o più variabili, la variabile somma segue la distribuzione normale, indipendentemente dalla distribuzione delle variabili originali (teorema del limite centrale).

2) La maggior parte delle variabili biologiche (peso, statura, …) dipendono dalla somma di svariati fattori genetici e ambientali.

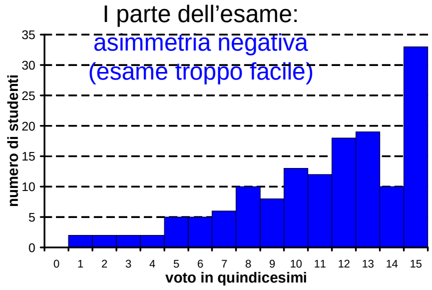
↓

La maggior parte delle variabili biologiche segue la distribuzione normale.

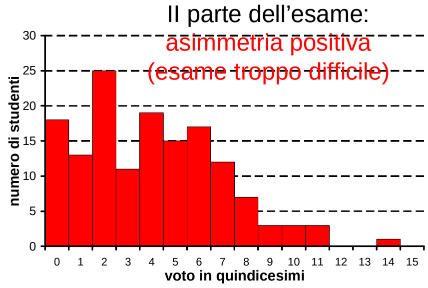
3) La maggior parte delle altre distribuzioni teoriche di probabilità (binomiale, di Poisson, t di Student) tendono alla distribuzione normale, all'aumentare della numerosità (o del numero di casi e di non-casi).

Verifica del 17 novembre 1994

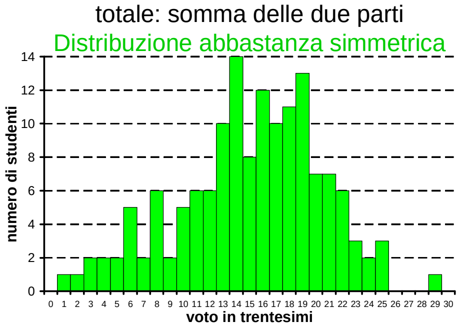
I parte dell'esame:



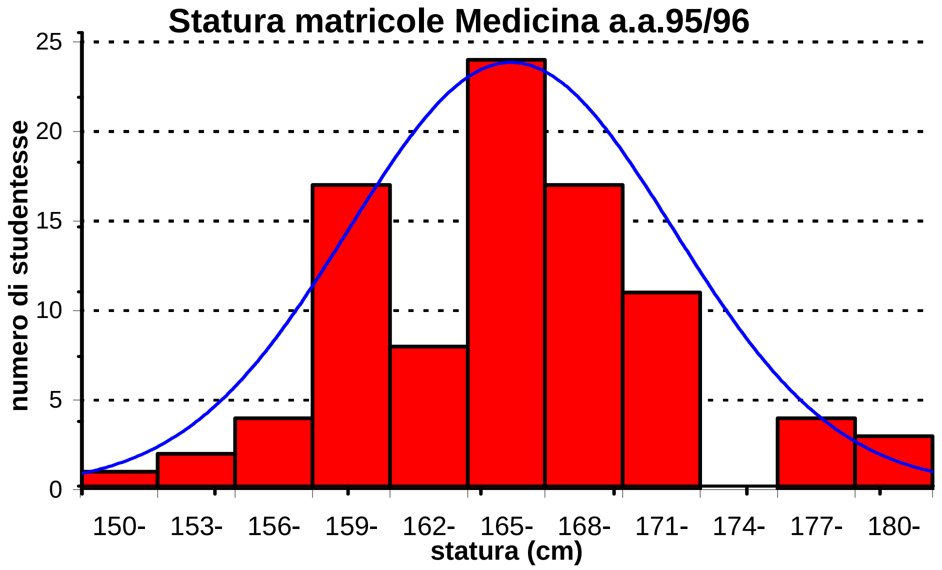
II parte dell'esame:



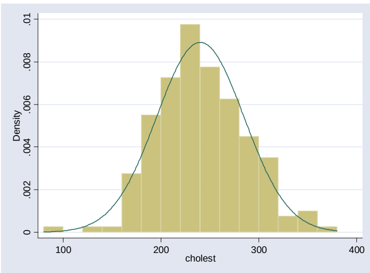
totale: somma delle due parti



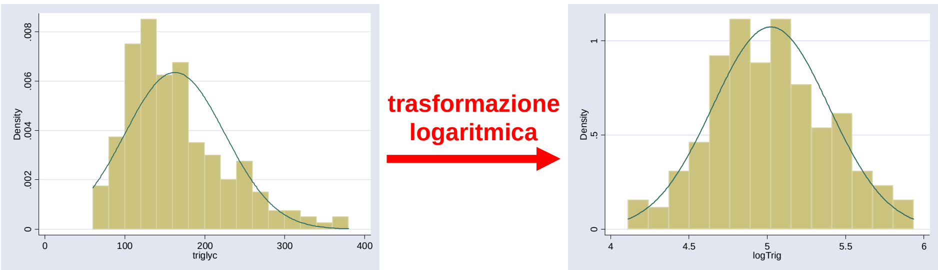
La distribuzione empirica (colonne rosse) può essere approssimata con una curva teorica (la distribuzione normale con media=166,1 cm e deviazione standard=6,1 cm).

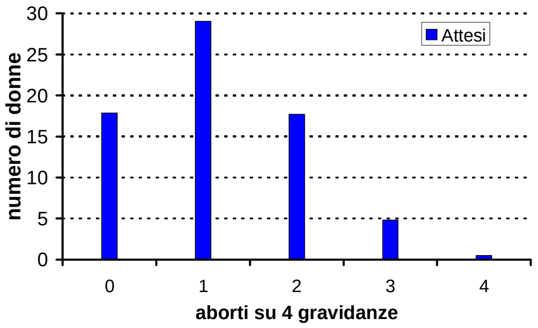


Distribuzione della colesterolemia in 200 diabetici di tipo 2



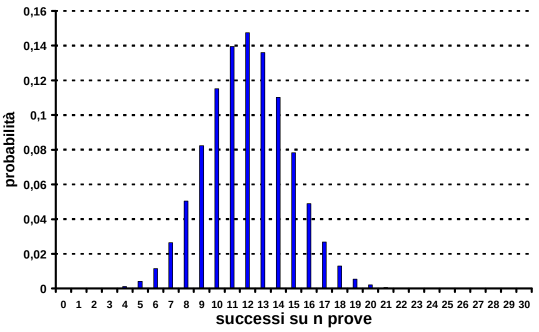
Distribuzione della trigliceridemia in 200 diabetici di tipo 2





Distribuzione binomiale con π=0.289 ed n=4.

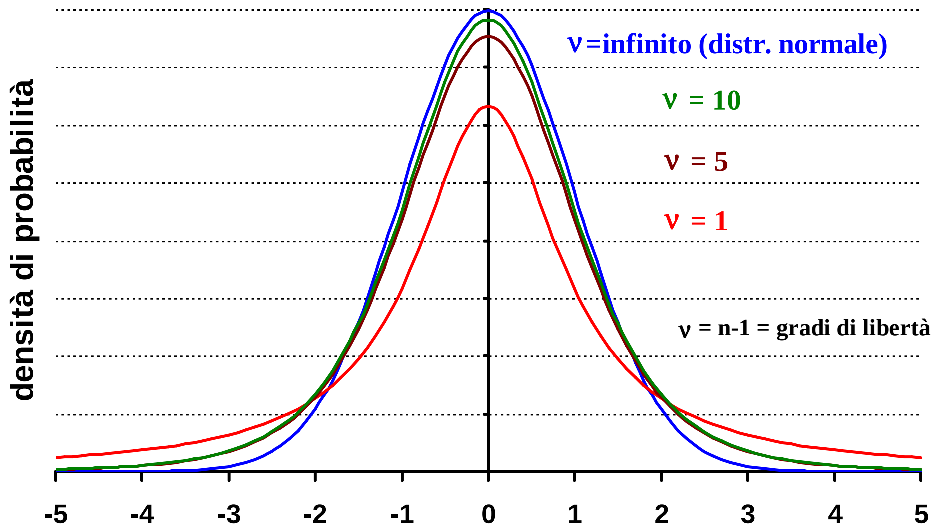
Distribuzione con forte asimmetria positiva.



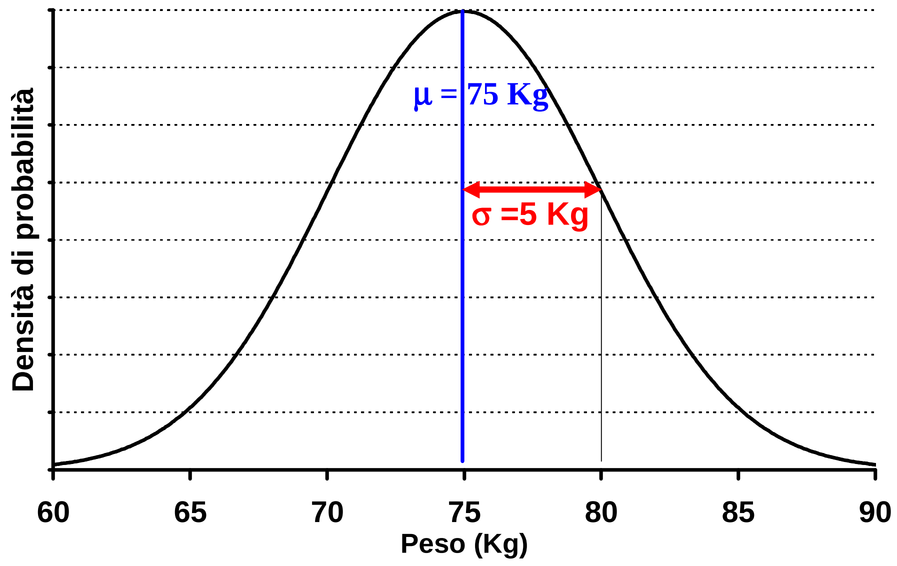
Distribuzione binomiale con π=0.4 ed n=30.

Distribuzione simmetrica, che tende alla distribu-zione normale.

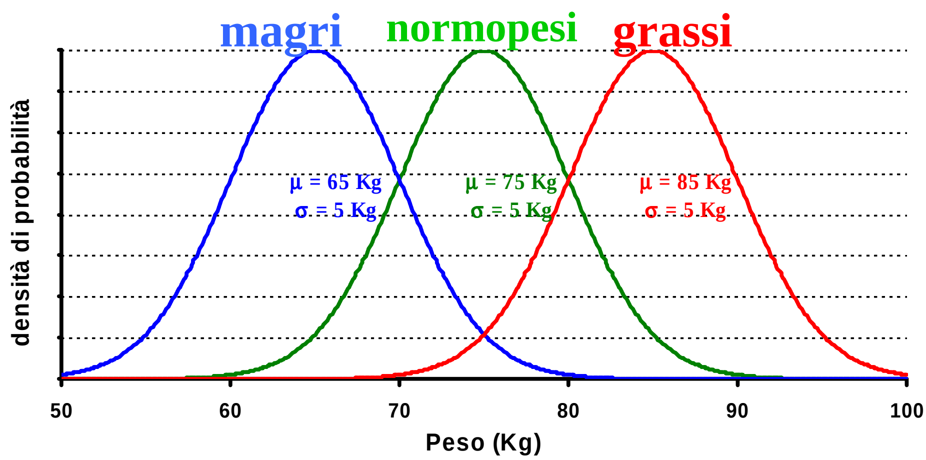
Distribuzione T di Student



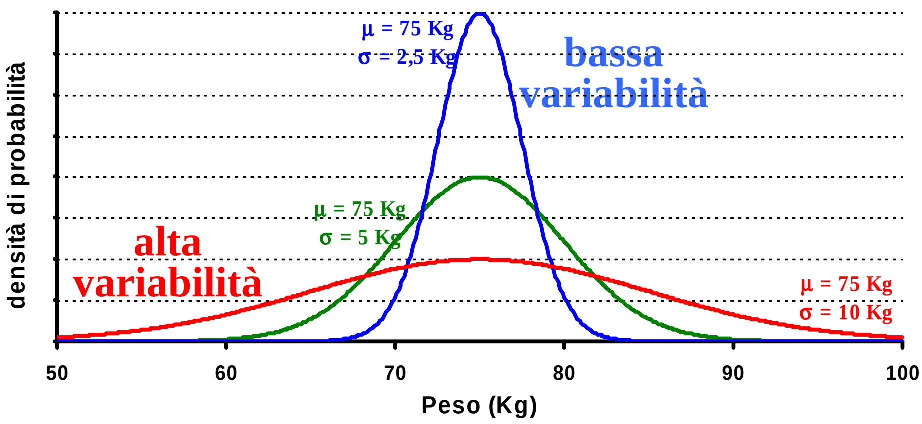
Una distribuzione normale è pienamente caratterizzata da due parametri, la media (μ) e la deviazione standard (σ)



Queste 3 distribuzioni differiscono per la media (misura di posizione)



Queste 3 distribuzioni differiscono per la deviazione standard (misura di dispersione)

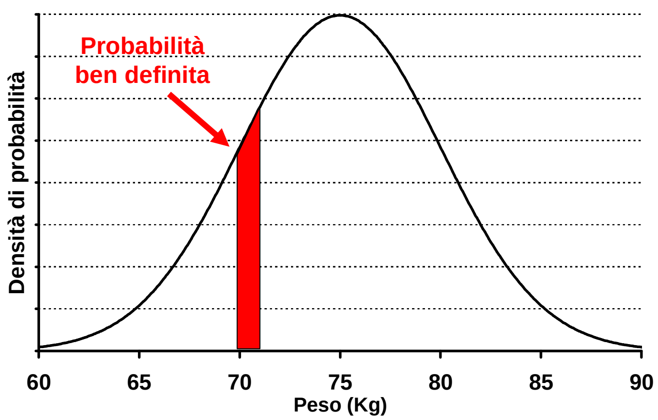


Probabilità e densità di probabilità

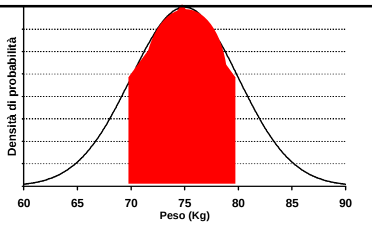
Il peso è una variabile quantitativa continua.

Qual è la probabilità che il peso di un individuo di questa popolazione sia esattamente 73Kg 133g 917mg 715µg 822ng? Praticamente zero.

Noi possiamo associare una probabilità non ad un singolo valore, ma ad un intervallo.

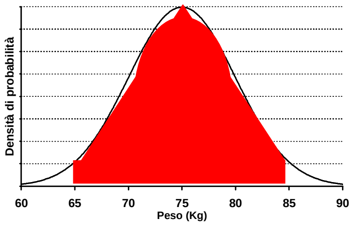


media ± 1 dev.standard =
75 ± 5 = 70-80 Kg



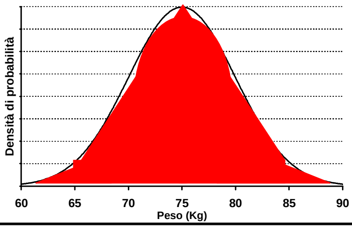
68,26%

media ± 2 dev.standard =
75 ± 10 = 65-85 Kg



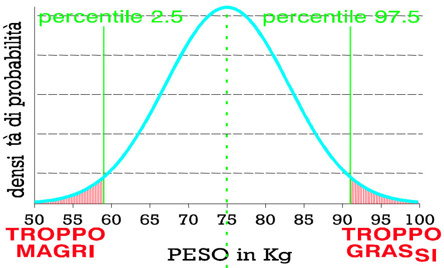
95,44%

media ± 3 dev.standard =
75 ± 15 = 60-90 Kg

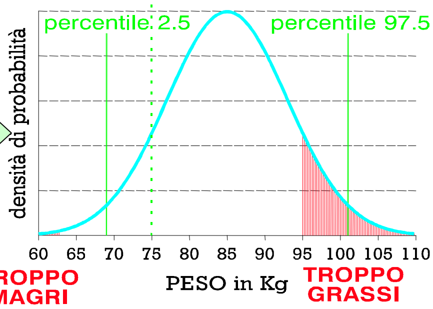


99,74%

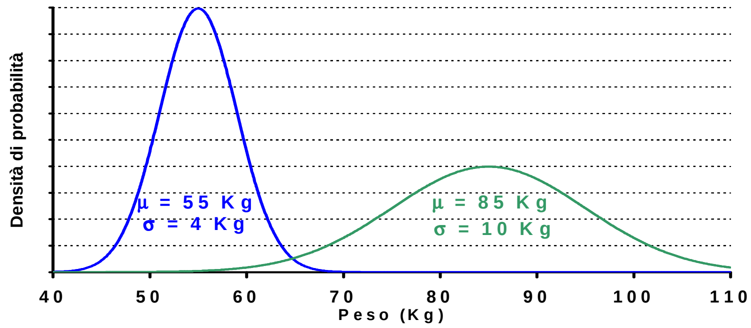
NORMALITA' STATISTICA



NORMALITA' CLINICA:
la popolazione è obesa nel suo complesso.
"The American population is constant in number, but it is ballooning in mass" (CDC, Atlanta, USA)



Esiste un numero infinito di distribuzioni normali diverse fra loro.

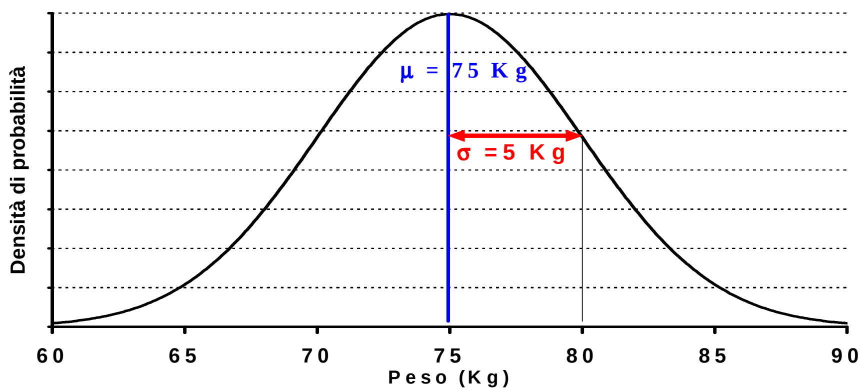


E' possibile ricondurre tutte queste diverse distribuzioni ad un'unica distribuzione standard?

Sì, attraverso la trasformazione normale:

z (deviata normale standardizzata) = $(x-\mu)/\sigma$

z (deviata normale standardizzata) = $(x-\mu)/\sigma$



z	-3	-2	-1	0	+1	+2	+3

Peso=60 Kg z = (60-75)/5 = -15/5 = -3

Peso=65 Kg z = (65-75)/5 = -10/5 = -2

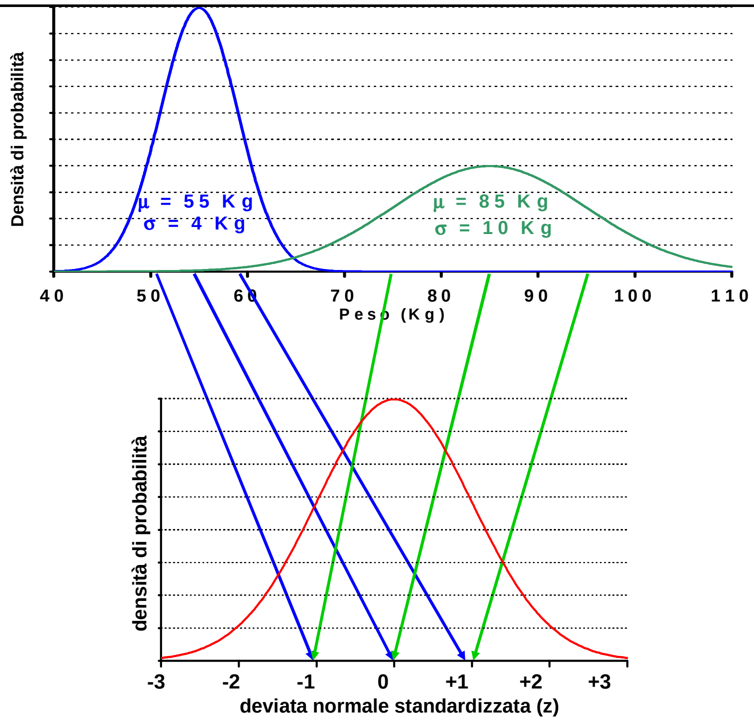
Peso=70 Kg z = (70-75)/5 = -5/5 = -1

Peso=75 Kg z = (75-75)/5 = 0/5 = 0

peso=80 Kg z = (80-75)/5 = +1

peso=85 Kg z = (85-75)/5 = +2

peso=90 Kg z = (90-75)/5 = +3



Esistono delle tavole (tavole della z) che danno la probabilità che Z sia maggiore di un valore qualsiasi.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,06430	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233

Qual è la probabilità che Z sia maggiore o uguale a 1,87?

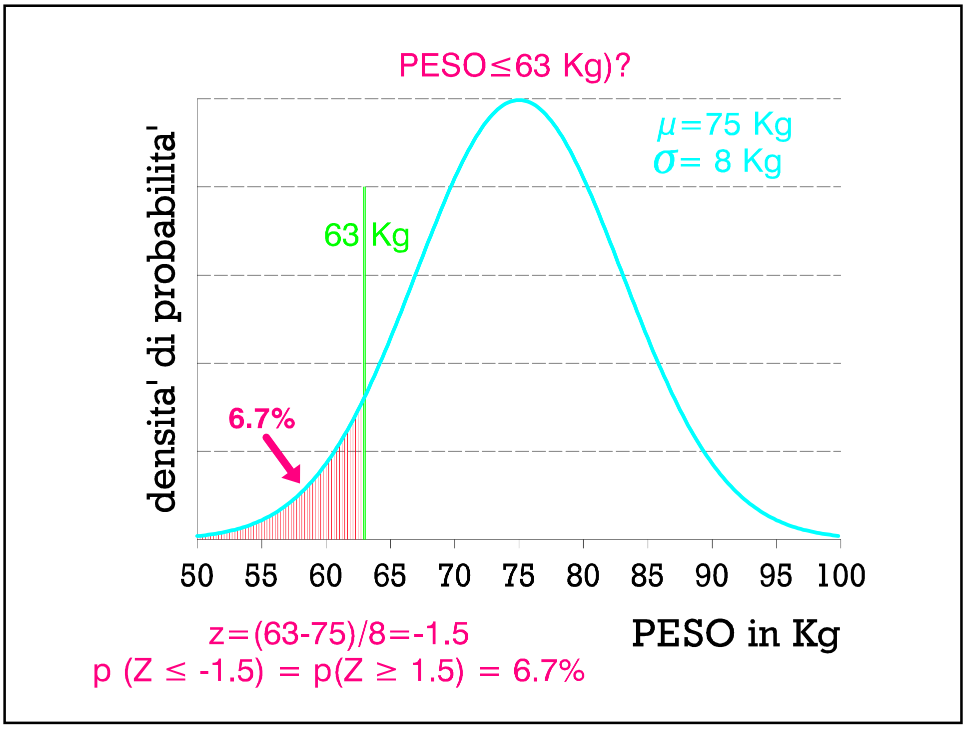
0,0307 = 3,07%

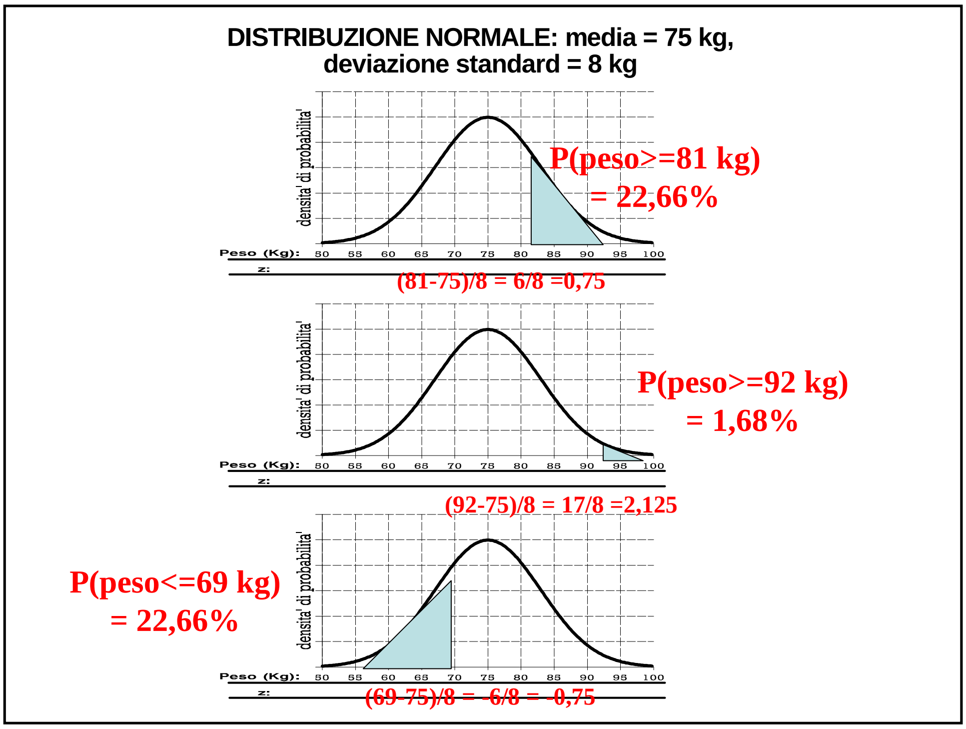
z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,06430	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233

Qual è la probabilità che Z sia maggiore o uguale a 0,75?

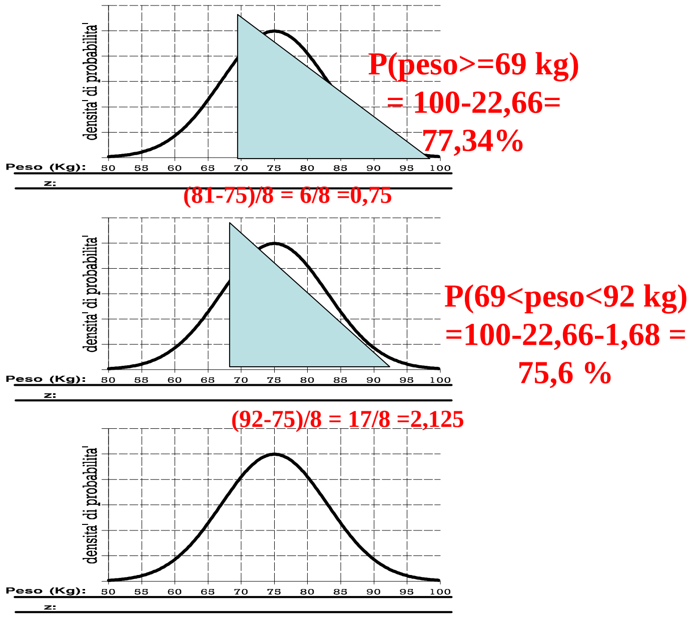
0,2266 = 22,66%

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,06430	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233

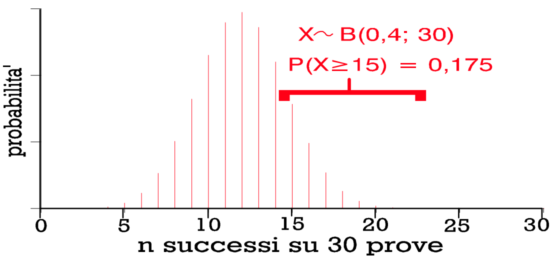

PESO≤63 Kg)?
μ=75 Kg
σ= 8 Kg
63 Kg
6.7%
densita' di probabilita'
50 55 60 65 70 75 80 85 90 95 100
PESO in Kg
z=(63-75)/8=-1.5
p (Z ≤ -1.5) = p(Z ≥ 1.5) = 6.7%


DISTRIBUZIONE NORMALE: media = 75 kg,
deviazione standard = 8 kg
densita' di probabilita'
P(peso>=81 kg)
= 22,66%
Peso (Kg): 50 55 60 65 70 75 80 85 90 95 100
z:
(81-75)/8 = 6/8 =0,75
P(peso>=92 kg)
= 1,68%
(92-75)/8 = 17/8 =2,125
P(peso<=69 kg)
= 22,66%
(69-75)/8 = -6/8 = -0,75

DISTRIBUZIONE NORMALE: media = 75 kg, deviazione standard = 8 kg



CORREZIONE PER LA CONTINUITA



z= (15-12)/2.68 =1,12 P(Z>=1,12)=0,131

z= (15-12-0,5)/2.68 =0,93 P(Z>=0,93)=0,176