Introduzione alle distribuzioni di probabilità di variabili continue

ESPERIMENTO

VARIABILE CASUALE
(V.C.)

caratteristica che può essere
DISCRETA o CONTINUA

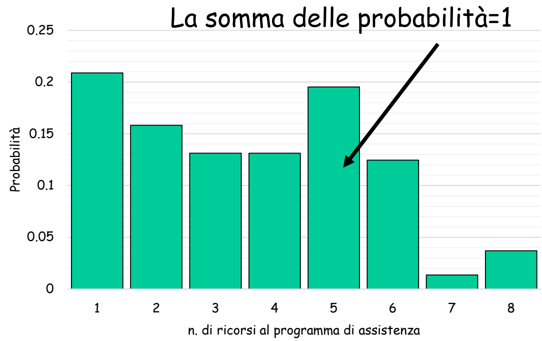
soggetta al caso

La distribuzione di probabilità di una v.c. discreta è ben definita da un prospetto, un grafico, una formula che consente di specificare tutti i possibili valori che la variabile assume con le rispettive probabilità

Variabile casuale discreta: esempio

N. di volte che le famiglie con bambini sono ricorse al programma di assistenza

N. volte	Freq.	P(X=x)
1	62	0.2088
2	47	0.1582
3	39	0.1313
4	39	0.1313
5	58	0.1953
6	37	0.1246
7	4	0.0135
8	11	0.0370
Totale	297	1.0000



$P(X=3)$?
$P(X \leq 2)$=?

DISTRIBUZIONE DI PROBABILITÀ

ESPERIMENTO

VARIABILE CASUALE
(V.C.)

caratteristica che può essere
DISCRETA o CONTINUA

soggetta al caso

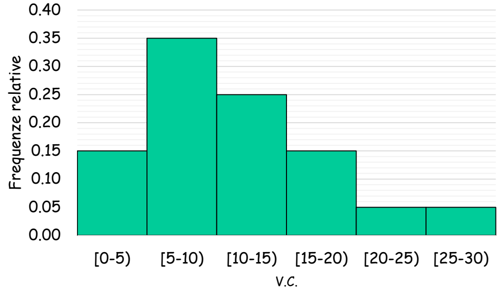
????????

Variabile casuale continua: esempio

Immaginiamo di raccogliere informazioni relative ad una V.C. continua che assume valori da 0 a 30. I dati che osserveremo saranno 0, 10, 22, 18, 7, 9, 2, 26, ...

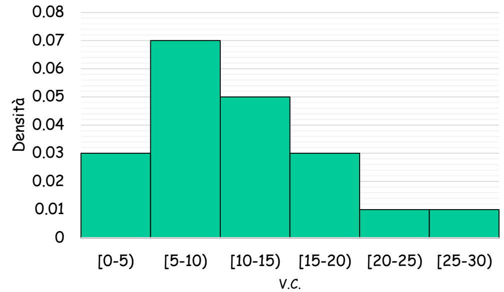
V.C.	$P(x_i<X<x_{i+5})$
[0-5)	0.15
[5-10)	0.35
[10-15)	0.25
[15-20)	0.15
[20-25)	0.05
[25-30)	0.05
Totale	1

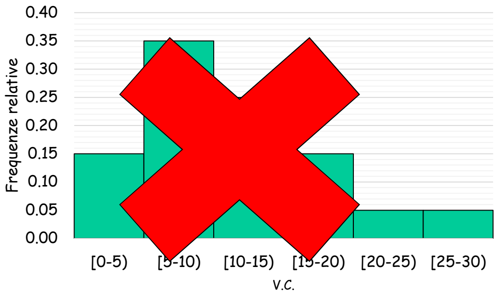
Come rappresentiamo la distribuzione di probabilità?



Qual è la distribuzione di probabilità corretta?

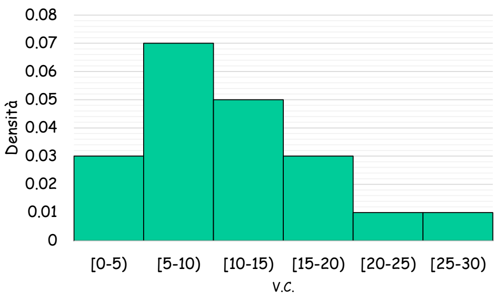
Ricordiamo che:
1) $0 \leq P(X=x) \leq 1$
2) $\sum P(X=x) = 1$

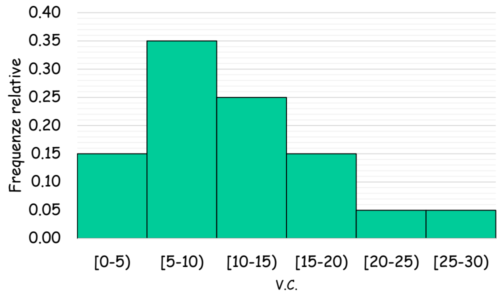




Qual è la distribuzione di probabilità corretta?

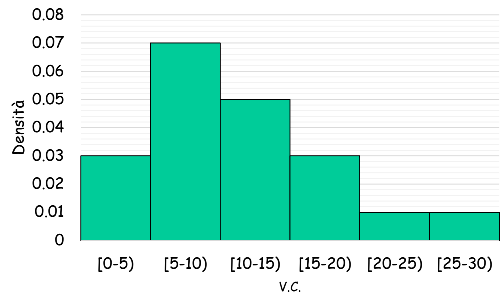
Ricordiamo che:
1) $0 \leq P(X=x) \leq 1$
2) $\sum P(X=x) = 1$





0.15x5	0.75
0.35x5	1.75
0.25x5	1.25
0.15x5	0.75
0.05x5	0.25
0.05x5	0.25
Totale	5

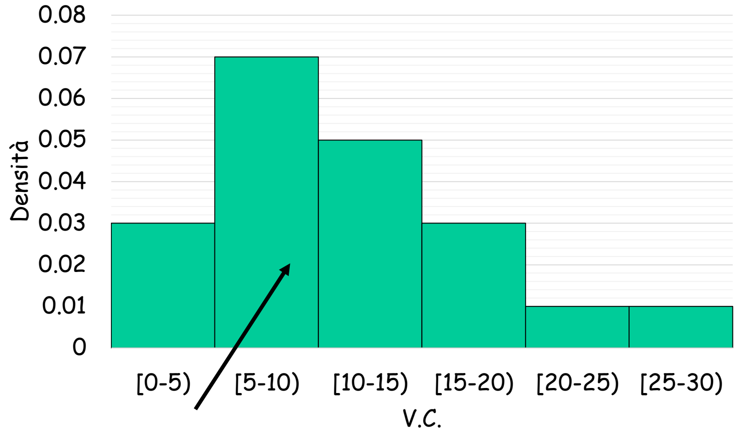
0.03x5	0.15
0.07x5	0.35
0.05x5	0.25
0.03x5	0.15
0.01x5	0.05
0.01x5	0.05
Totale	1



Istogramma

Altezza rettangolo (**DENSITÀ**) =

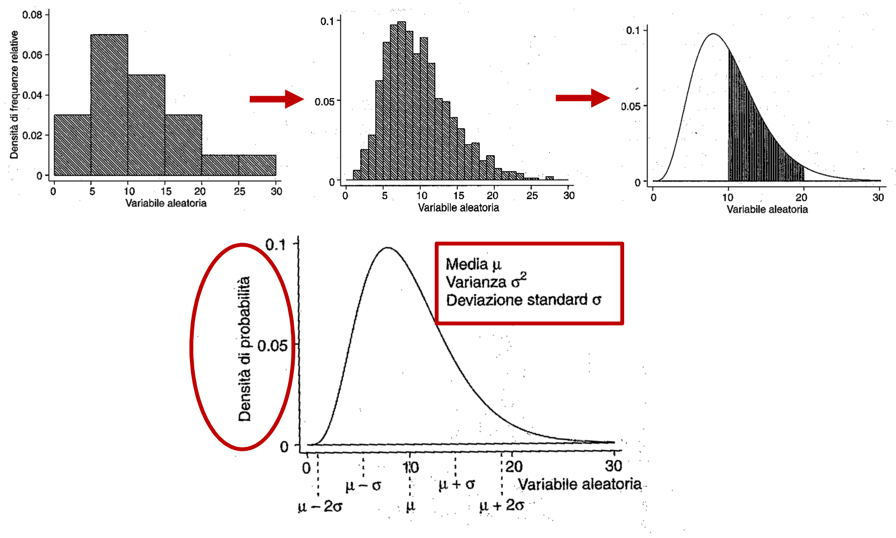
= frequenza relativa/ampiezza della classe



Area totale sottesa = 1

Dal discreto al continuo…

… classi sempre più piccole, n ->∞



Variabile casuale continua

Può assumere un numero infinito di valori compreso in un intervallo di ampiezza finita o infinita.

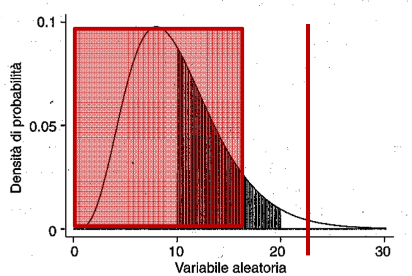
->La probabilità per ogni singolo valore è pari a 0 $P(X=x)=0$

->Si assegna una probabilità per un intervallo di valori $P(a<X<b)\geq 0$

Esempio

Qual è la probabilità di avere un BMI di 23 kg/m^2?

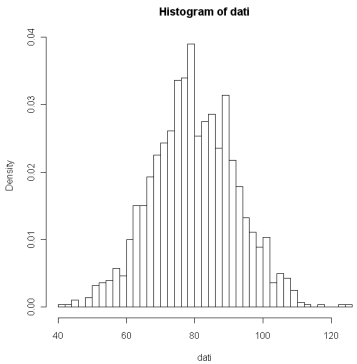
Qual è la probabilità di avere un BMI < 18 kg/m^2?



GAUSSIANA

Esempio: Pressione diastolica

La PAD è stata misurata in un campione di 1500 uomini tra i 35 e 44 anni. I risultati sono rappresentati con un istogramma delle frequenze relative divise per ampiezza della classe di PAD (classi di 2 mmHg)

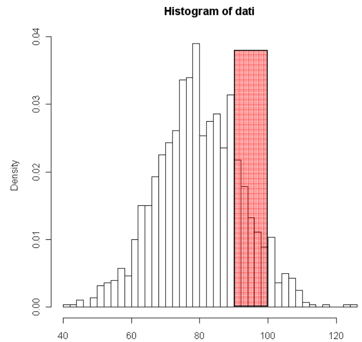


$\bar{x} = 80$

$s = 12$

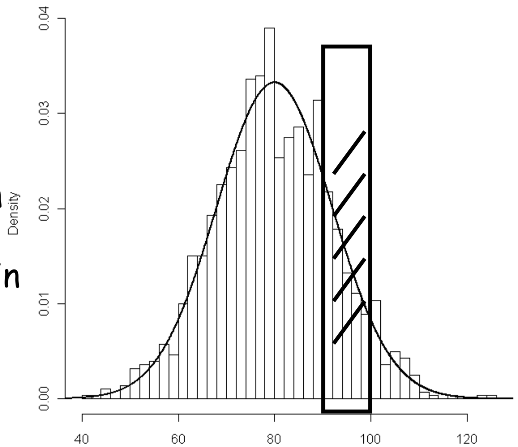
Usando l'istogramma potrei anche trovare la % (probabilità) di avere PAD in intervalli di interesse clinico

- ad esempio **90 , 100**
- dovrei sommare le aree dei rettangoli che 'coprono' l'intervallo in questione …



Può diventare un'operazione abbastanza laboriosa

L'istogramma è una funzione a gradini che in tutti quei casi dove presenta simmetria, ed in generale per molti fenomeni naturali/sperimentali,



- per n-> ∞ si può approssimare con una funzione continua appartenente alla famiglia della distribuzione Gaussiana (Normale)
- **le aree sottese si ottengono mediante <u>integrali</u>**

La funzione di densità di probabilità (f.d.p) è una funzione continua

Proprietà:

1. $f(x) \geq 0$ per ogni x
2. $\int_{-\infty}^{+\infty} f(x)dx = 1$
3. $P(a \leq x \leq b) = \int_a^b f(x)dx$

Funzione di distribuzione F (X)

$$F(X) = \int_{-\infty}^{x} f(t)dt$$

Valore atteso e varianza di una v.c. continua

Anche per le v.c. continue possiamo definire il valore atteso

$$E(X) = \int_{\Omega} x f(x) dx = \boldsymbol{\mu}$$

e la varianza

$$\text{Var}(X) = \int_{\Omega} [x - E(X)]^2 f(x) dx = \boldsymbol{\sigma^2}$$

Importanza della V.C. Gaussiana

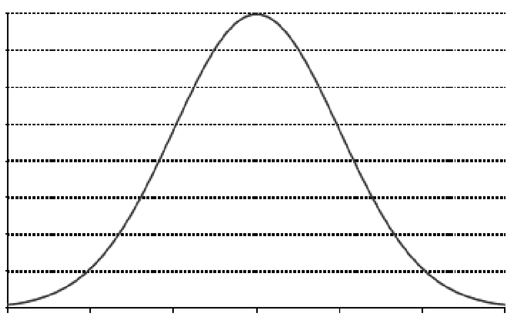
La v.c. Gaussiana riveste un ruolo fondamentale perché:

- **descrive bene** il manifestarsi di **molti fenomeni**, per esempio:
  - Errori di misura (genesi della Gaussiana)
  - Caratteristiche morfologiche (altezza, lunghezza)
- gode di **importanti proprietà** (aspetto tecnico rilevante)

Distribuzione normale

Detta anche **gaussiana** dal nome di Karl Friedrich Gauss (1777-1855).

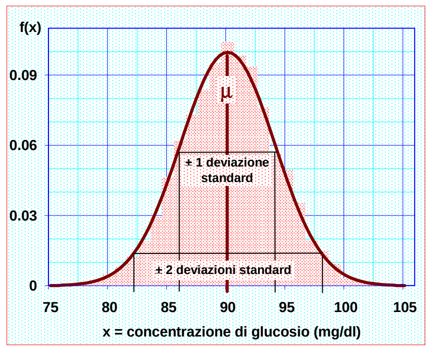
Occupa un ruolo centrale nella statistica



Errori di misura

Gli errori casuali di misura ($\varepsilon = x - \mu$), considerati nel loro complesso, mostrano un comportamento tipico che può essere così descritto:

1. gli ***errori piccoli*** sono più frequenti di quelli ***grandi***;

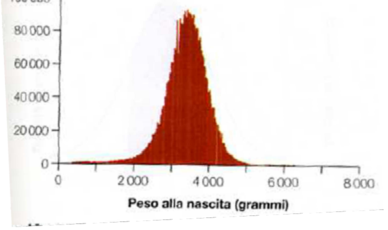


2. gli errori di ***segno negativo*** tendono a manifestarsi con la stessa frequenza di quelli con segno positivo;
3. all'aumentare del numero delle misure si ha che:

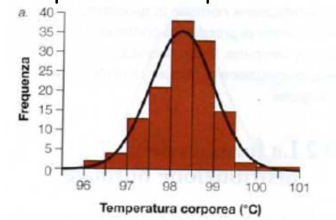
- i **2/3** dei valori tendono ad essere inclusi nell'intervallo **media±1 deviazione standard**
- il **95% dei valori tende** ad essere incluso nell'intervallo **media±2 deviazioni standard**

Esempi di variabili con distribuzione simmetrica

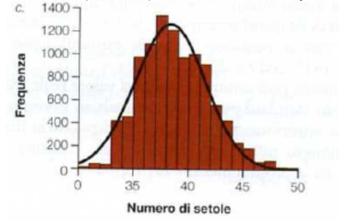
Peso alla nascita



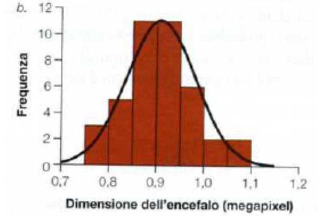
Temperatura corporea umana



Numero di setole sul 4° e 5° segmento dell'addome in moscerini della frutta

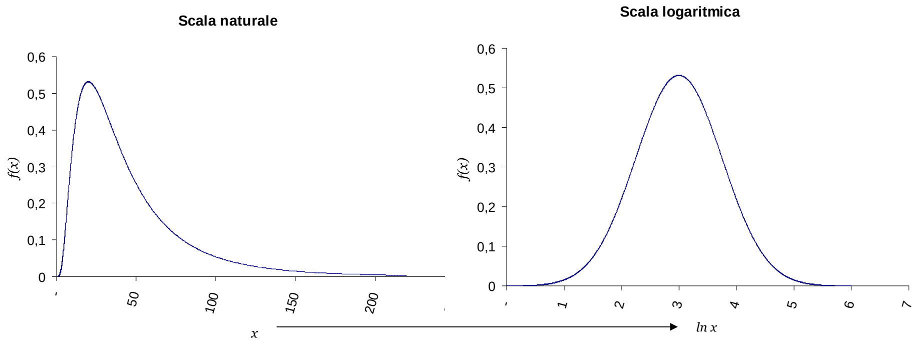


Dimensione dell'encefalo degli studenti universitari



Variabili da trasformare

Esempio: trigliceridi nel siero

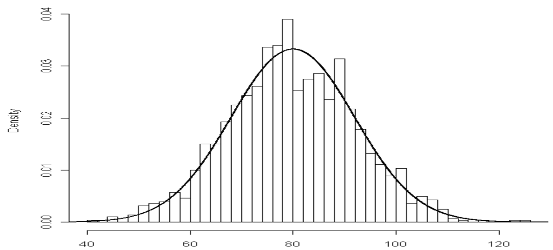


Distribuzione Gaussiana

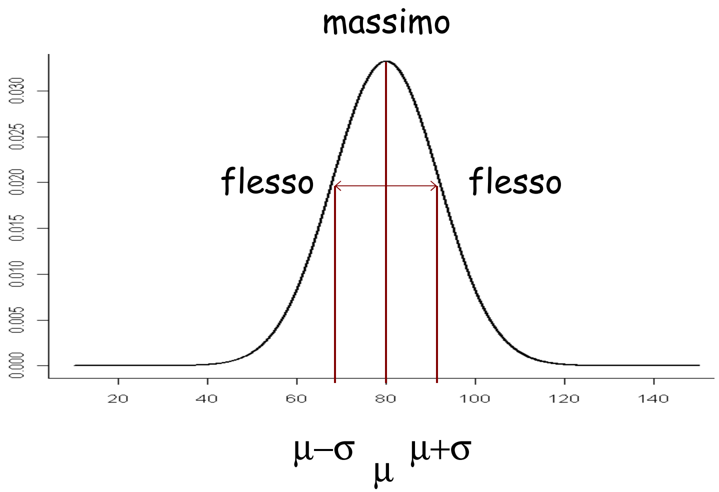
La famiglia di funzioni Gaussiane dipende da due parametri μ e σ (>0)

$$y = f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma} \right)^2 \right]$$

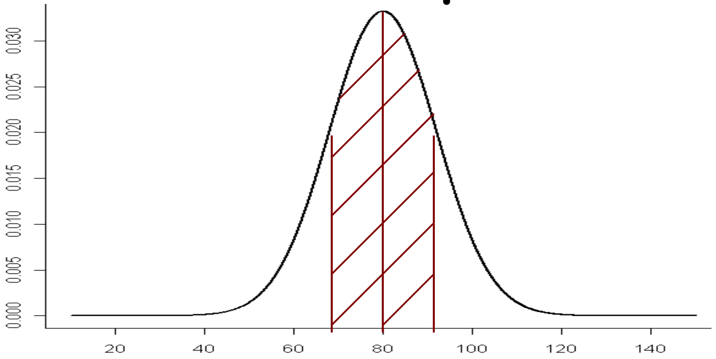
Se si pone $\mu = \bar{x}$ e $\sigma = s$ si ottiene una buona approssimazione



Forma funzionale



Intervalli tipo

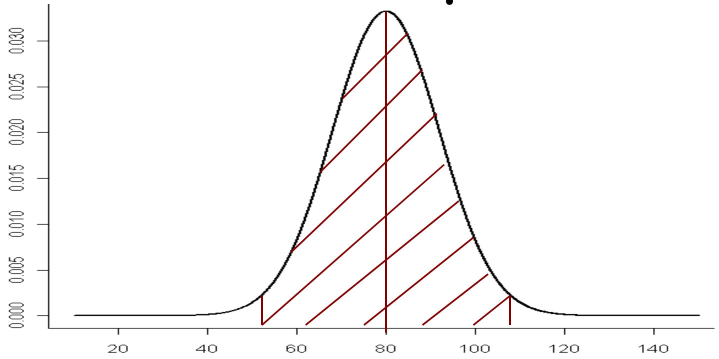


$\mu-\sigma$ μ $\mu+\sigma$
$(\mu-1\sigma)$ $(\mu+1\sigma)$

Area sottesa = 0.68 = 68%

Es. Il 68% della popolazione ha un valore di pressione compreso tra 80-12=68 mmHg e 80+12=92 mmHg

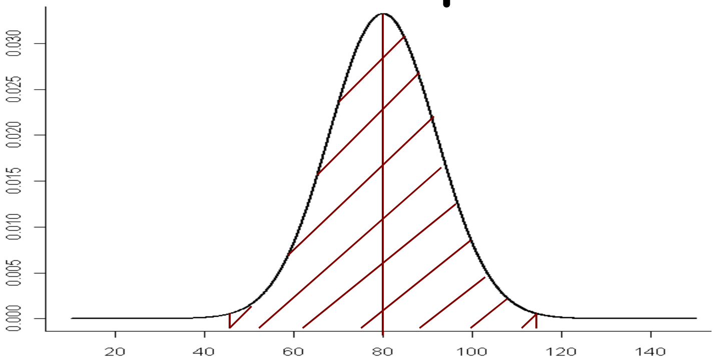
Intervalli tipo



$\mu-1.96\sigma$ μ $\mu+1.96\sigma$
$(\mu-2\sigma)$ $(\mu+2\sigma)$

Area sottesa = 0.95 = 95%

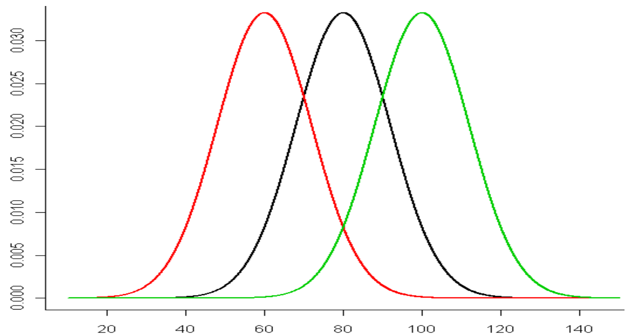
Intervalli tipo



$\mu-2.58\sigma$ μ $\mu+2.58\sigma$

Area sottesa = 0.99 = 99%

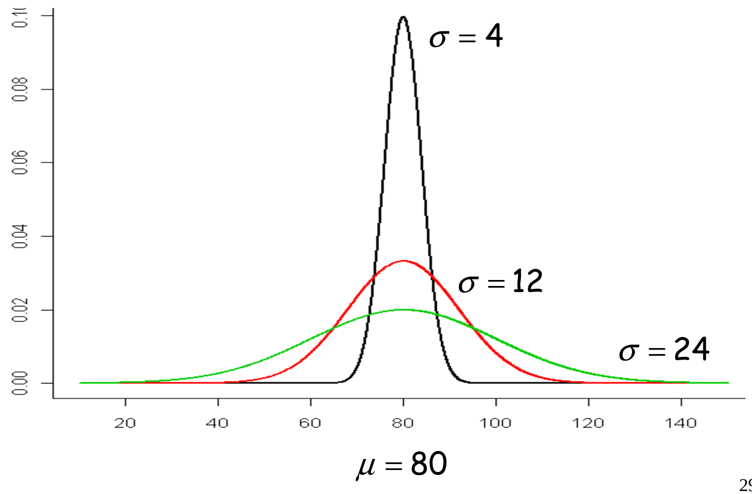
Dipendenza dai parametri



$\mu = 60$ $\mu = 80$ $\mu = 100$

$\sigma = 12$ $\sigma = 12$ $\sigma = 12$

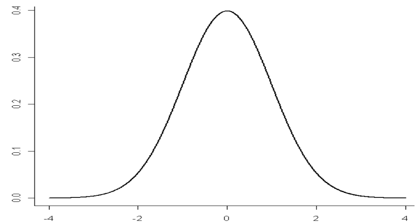
Dipendenza dai parametri



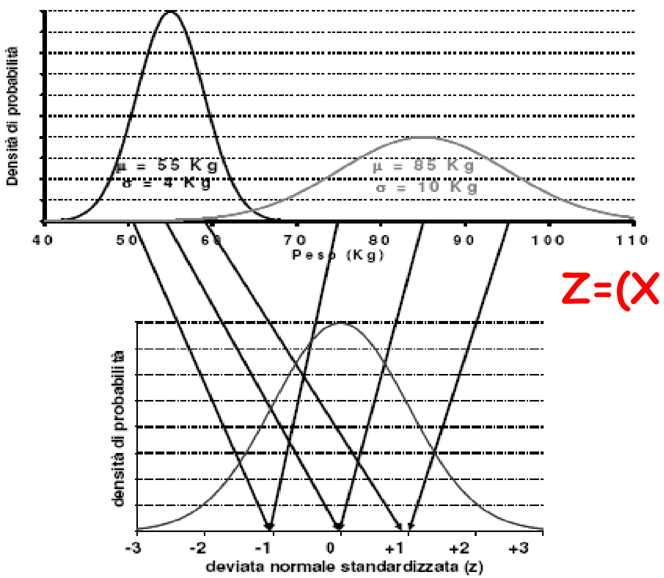
Gaussiana 'particolare'- Standardizzata

Funzione Gaussiana con parametri $\mu = 0$ e $\sigma = 1$

$$y = f(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} \cdot x^2\right]$$



Le funzioni gaussiane **non sono integrabili** ed andrebbero tabulate.
E' possibile però esprimere l'area sottesa da una generica gaussiana in termini di gaussiana standardizzata
E' stata quindi tabulata SOLO la gaussiana standardizzata
NORMALE (0,1)



Z=(X-μ)/σ

Tabella A1. Aree in una coda della curva normale standardizzata

Questa tabella riporta l'area tratteggiata

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	0.0643	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
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010

Lettura della tabella della normale standardizzata

Tabella A1. Aree in una coda della curva normale standardizzata

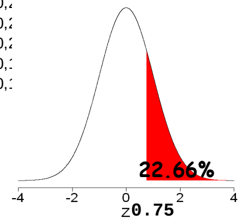
Questa tabella riporta l'area tratteggiata

2°decimale

Unità e 1°decimale

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	0,2		
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2		
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2		
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1		
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1		

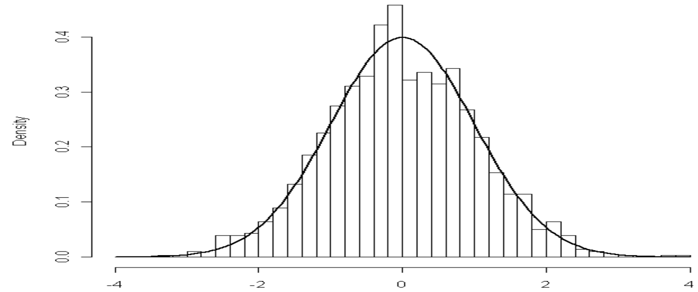
Es. **Pr(Z≥0.75)=0.2266**



Tornando all'istogramma della pressione diastolica approssimabile con gaussiana. Se ai valori *X* di *PAD* applichiamo questa trasformazione

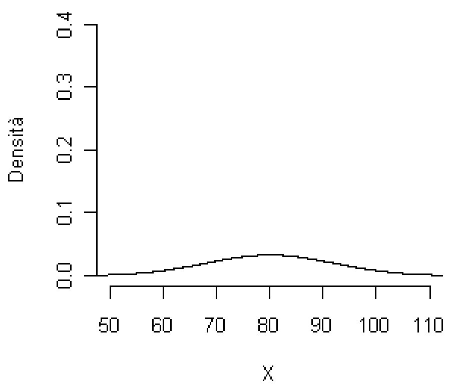
$$Z = \frac{X - 80}{12}$$

otteniamo un istogramma approssimabile con una gaussiana standardizzata



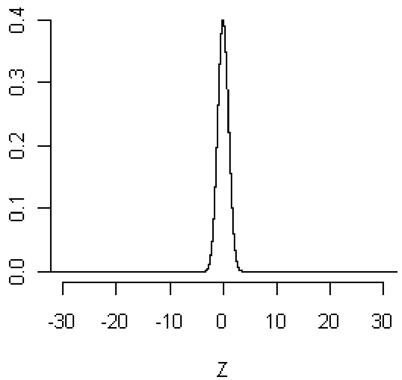
Quindi le aree sottese dalla Gaussiana

$\mu = 80$
$\sigma = 12$



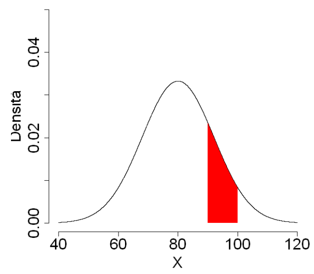
corrispondono ad aree sottese dalla Gaussiana standardizzata

$\mu = 0$
$\sigma = 1$



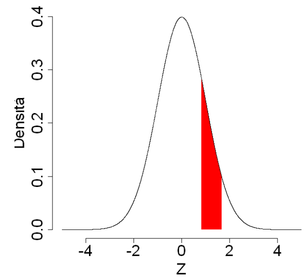
L'area tra 90 e 100

$\mu = 80$
$\sigma = 12$



corrisponde all'area tra (90-80)/12=0.83 e (100-80)/12=1.67

$\mu = 0$
$\sigma = 1$



Tali aree che corrispondono a delle <u>frequenze relative</u> in termini descrittivi, vengono pensate come delle <u>probabilità</u> in termini predittivi.

Riscrivo quindi la mia domanda originale come probabilità

$$\Pr\{90 \le X < 100\}$$

In termini di Gaussiana standardizzata

$$\Pr\left\{\frac{90-80}{12} \le \frac{X-\mu}{\sigma} < \frac{100-80}{12}\right\} = \Pr\{0.83 \le Z < 1.67\}$$

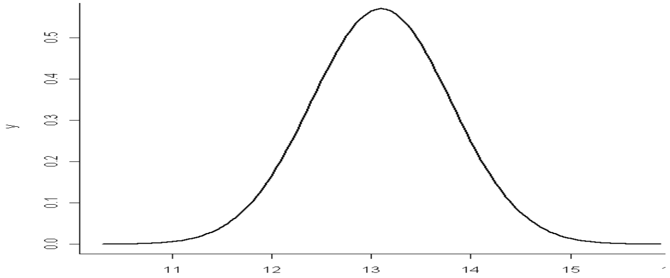
che posso calcolare in termini di Gaussiana standardizzata

$$\Pr\{Z > 0.83\} = 0.20 \qquad \Pr\{Z > 1.67\} = 0.05$$

$$\Pr\{0.83 \le Z < 1.67\} = 0.20 - 0.05 = 0.15 = 15\%$$

Esercizio per lo studente

In una popolazione di ragazze di età inclusa tra i 18 e i 25 anni, la concentrazione di emoglobina nel sangue (X) si approssima con una gaussiana con media μ = 13.1 g/dl e deviazione standard σ = 0.7 g/dl.

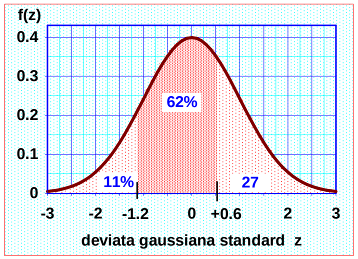


- Che percentuale di ragazze hanno emoglobinemia inclusa tra 12.26 e 13.52 g/dl ?

… continua

z_1 = (12.26-13.10) / 0.7 = -1.2

z_2 = (13.52-13.10) / 0.7 = +0.6



Nell'11% delle ragazze i valori di Hb sono minori di 12.26 g/dl, e nel 27% sono maggiori di 13.52 g/dl.
Quindi il 62% delle ragazze ha valori di Hb compresi tra 12.26 e 13.52 g/dl.

Esercizio per lo studente

In generale nella popolazione il livello di colesterolo plasmatico è distribuito normalmente con media 219 mg/dL e deviazione standard 50 mg/dL

1. Il livello *desiderabile* di colesterolo è <200 mg/dL. Quale % di persone ha un livello di colesterolo desiderabile ?
2. Un livello di colesterolo > 250 mg/dL sembra correlato ad un rischio sufficiente da consigliare il trattamento. Quale % di persone ha bisogno di trattamento ?
3. Calcolare senza effettuare conti la % di persone con colesterolo maggiore di 319
4. Qual è il valore di colesterolo oltre il quale si trovano il 10% dei soggetti con colesterolo più alto?

Malattie cardiovascolari e colesterolo plasmatico

Standardizziamo

1. Per x=200 ottengo

$$z = \frac{x-\mu}{\sigma} = \frac{200-219}{50} = -0,38$$

2. Per x=250 ottengo

$$z = \frac{x-\mu}{\sigma} = \frac{250-219}{50} = 0,62$$

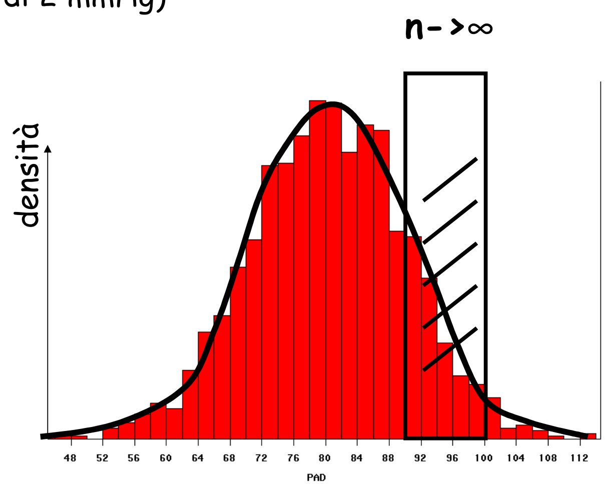
Malattie cardiovascolari e colesterolo plasmatico

1. Dalla tabella 1:
   - **Pr(*z*<-0,38)=0,352o**
   - Il **35,2%** della popolazione ha un livello di colesterolo desiderabile
2. Dalla tabella 1:
   - **Pr(*z*>0,62)=0,2676**
   - Il **26,8%** della popolazione potrebbe aver bisogno di un trattamento per l'ipercolesterolemia

3. (319-219) / 50 = 2 ➡ 2.5%

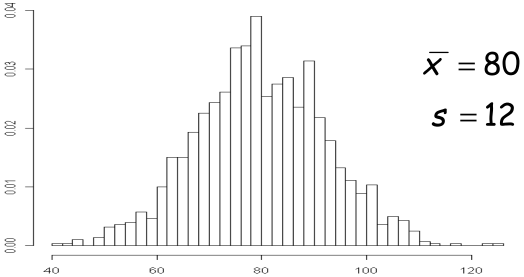
4. 1.28 = (x – 219) / 50
   x = 283

Se ad esempio misurassi la pressione arteriosa diastolica (PAD) in un campione di 1500 uomini tra i 35 e 44 anni potrei rappresentare i risultati in un *istogramma* delle frequenze relative divise per ampiezza della classe di PAD (classi di 2 mmHg)



Esempio: Pressione diastolica

La PAD è stata misurata in un campione di 1500 uomini tra i 35 e 44 anni. I risultati sono rappresentati con un istogramma delle frequenze relative divise per ampiezza della classe di PAD (classi di 2 mmHg)



$\bar{x} = 80$

$s = 12$

1) Il 68% delle persone hanno PAD tra 80-12 e 80+12
2) Il 95% delle persone hanno PAD tra 80-2*12 e 80+2*12
3) Il 99% delle persone hanno PAD tra 80-3*12 e 80+3*12