La distribuzione gaussiana

Caratteristiche della v.c. normale

- E' un modello teorico di distribuzione di una variabile casuale.
- Si applica alla probabilità ma anche alla distribuzione dei punteggi di un test o di altra misurazione
- È simmetrica intorno alla media.
- Media, moda e mediana coincidono.
- I quartili (25°,50°, e 75° centile) sono individuabili con le tavole della curva normale standard.

• Definizione analitica della curva normale

Formula della curva normale
(densità e ripartizione)

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

$\sigma =$ Dev stand

$\mu =$ Media

$\pi =$ pi greco

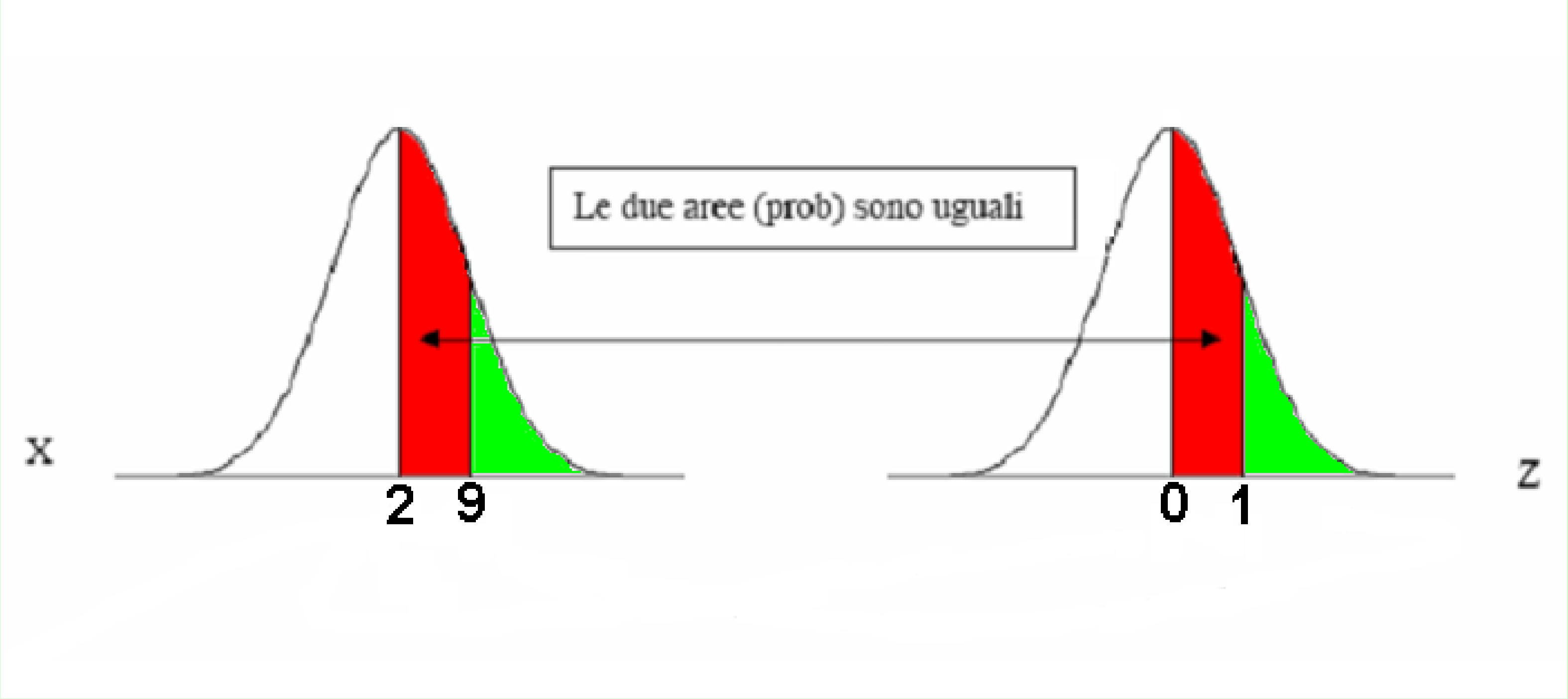
$e =$ 2,7183

$$F(X) = \int_{-\infty}^{X} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

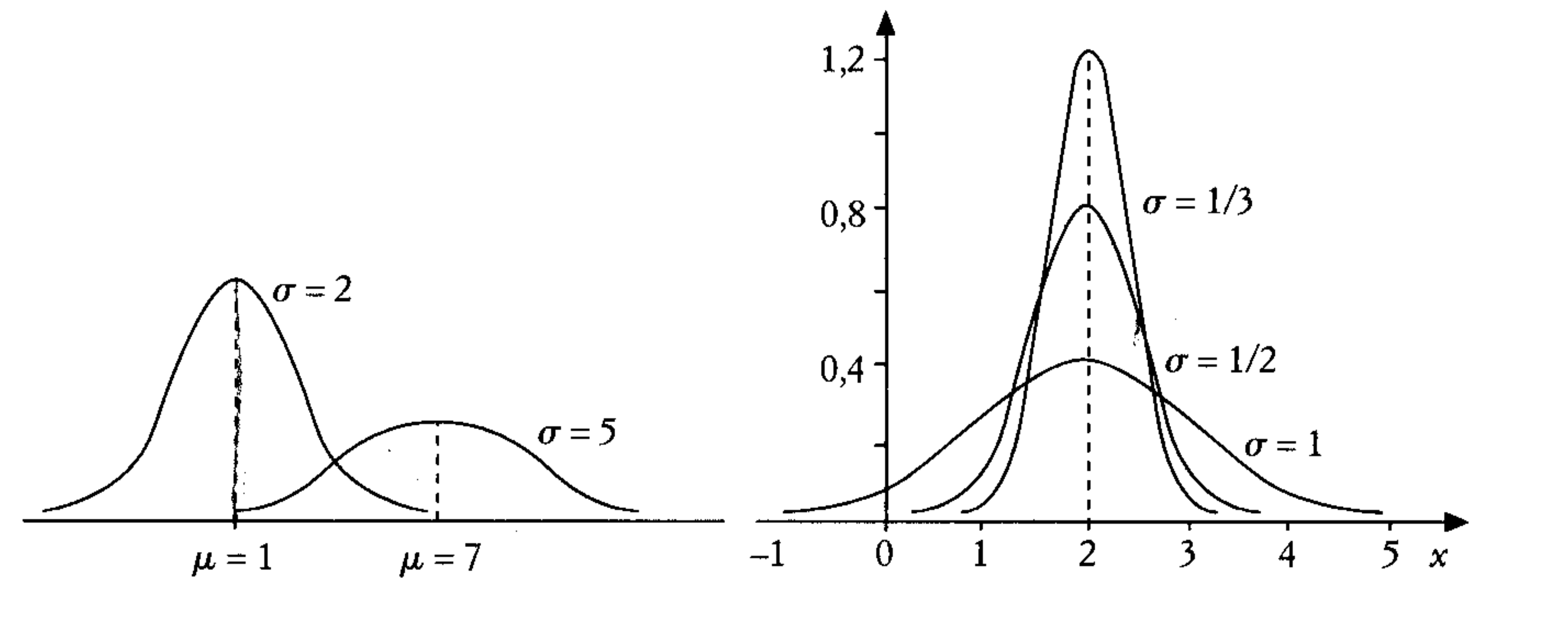
Normale standard

- la normale standard (riferibile alle tavole) con media = 0 e ds = 1 è quella di riferimento
- L'area totale dello spazio al di sotto della curva normale è uguale a 1 (oppure al 100% dei casi).
- Considerando solo una parte di quest'area si può calcolare, ad esempio, la percentuale dei casi che ha ottenuto un certo punteggio in un test o altra misurazione.

Due curve normali, anche se con parametri diversi, sono simili,



Due curve normali, anche se con parametri diversi, sono simili,



Standardizzare una distribuzione normale

La standardizzazione di una variabile permette il ricorso alla curva normale standard

Per **standardizzare** una distribuzione e trasformare i punteggi in punteggi standardizzati, è necessario eseguire l'operazione:

$$z_i = \frac{x_i - \mu}{\sigma}$$

EQUAZIONE ASSOCIATA ALLA DISTRIBUZIONE NORMALE STANDARDIZZATA:

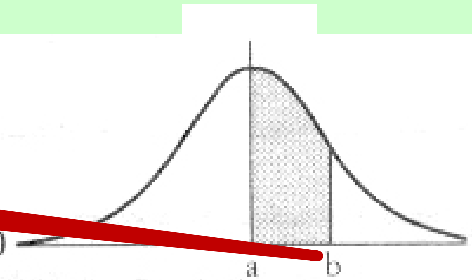
N(0,1)=z ⇨ $$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Come si usano le tavole della normale standard

Servono per individuare la percentuale di casi al disopra (o al disotto o compresi fra due valori)

TAVOLA

Troviamo un certo punto, per esempio b = 0,52

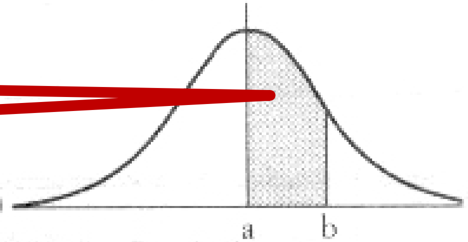


TAV. B. Aree della distribuzione normale standard tra $a = 0$ e $b > 0$

z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4351	.4265	.4279	.4292	.4306	.4319

TAVOLA

Al punto b = 0,52 corrisponde l'area 0,1985

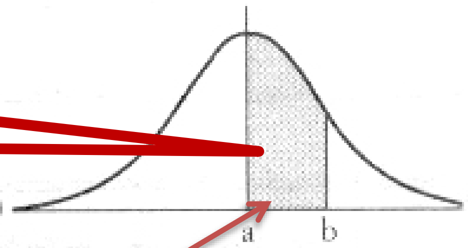


TAV. B. Aree della distribuzione normale standard tra $a = 0$ e $b > 0$

z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4351	.4265	.4279	.4292	.4306	.4319

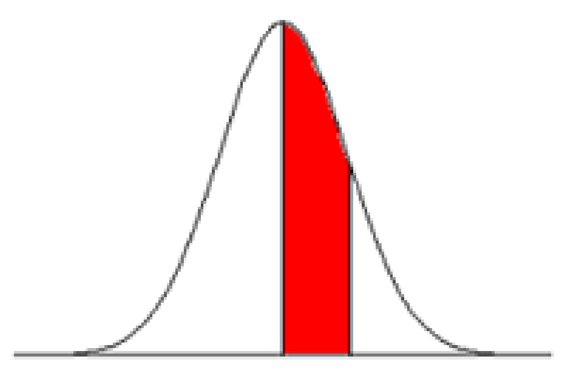
TAVOLA

All'area in grigio (0,1985) va sommato la metà sinistra (0,5) e il risultato dà 0,6985, pari a 69,85 %



TAV. B. Aree della distribuzione normale standard tra $a = 0$ e $b > 0$

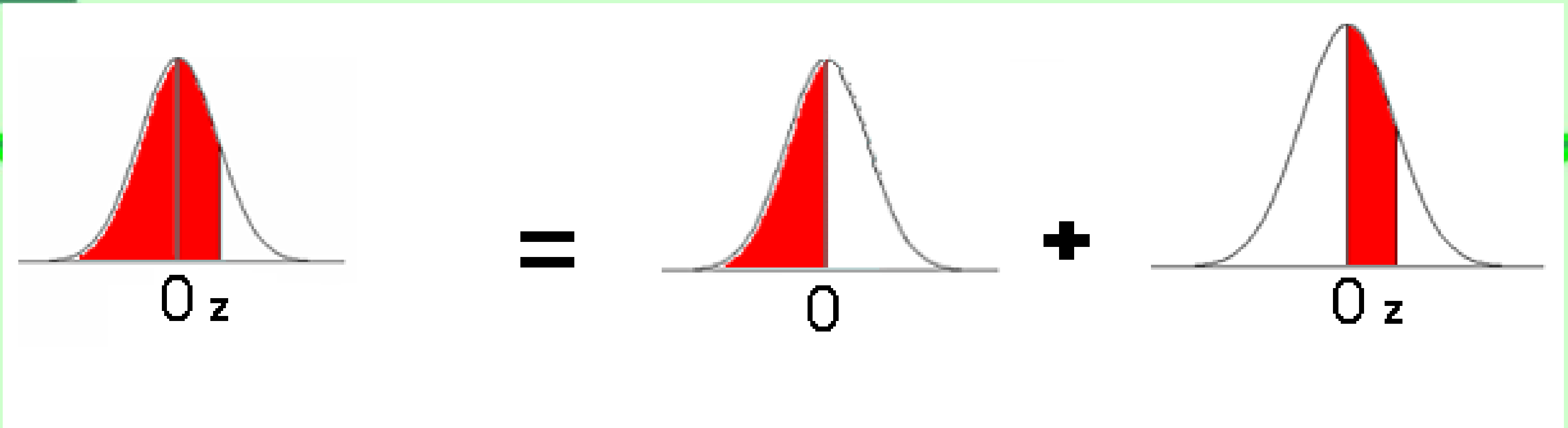
z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	$\boxed{.1985}$	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4351	.4265	.4279	.4292	.4306	.4319



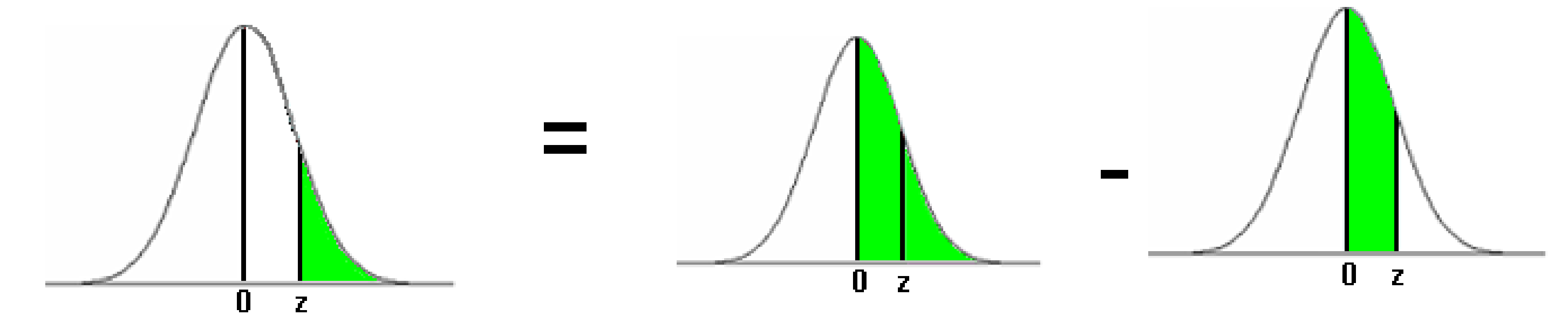
- I valori trovati nelle tavole corrispondono alle AREE comprese tra lo 0 (media della distribuzione) e il punto z da cui si è partiti (area rossa del disegno).

- Per trovare le aree corrispondenti ad altri intervalli (p.e. l'area dal punto z alla fine della curva) è necessario procedere con ulteriori calcoli ricordando che:

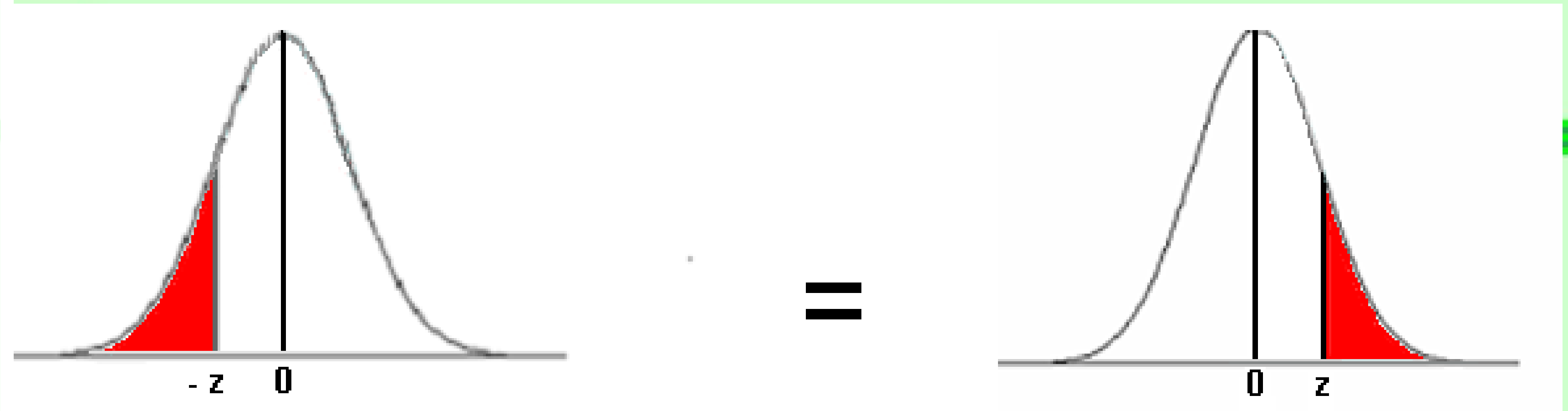
✓ L'area sottesa a tutta la curva è pari a 1
✓ La curva è simmetrica rispetto alla media 0
✓ La media divide l'area sotto la curva in due parti uguali, 0,5 a destra e 0,5 a sinistra



Applicando questo calcolo ai punteggi di un test, la probabilità P che i soggetti abbiano realizzato un punteggio standardizzato X **minore** di un certo valore z si trova sommando a 0.5 la probabilità che il loro punteggio stia tra 0 e z.

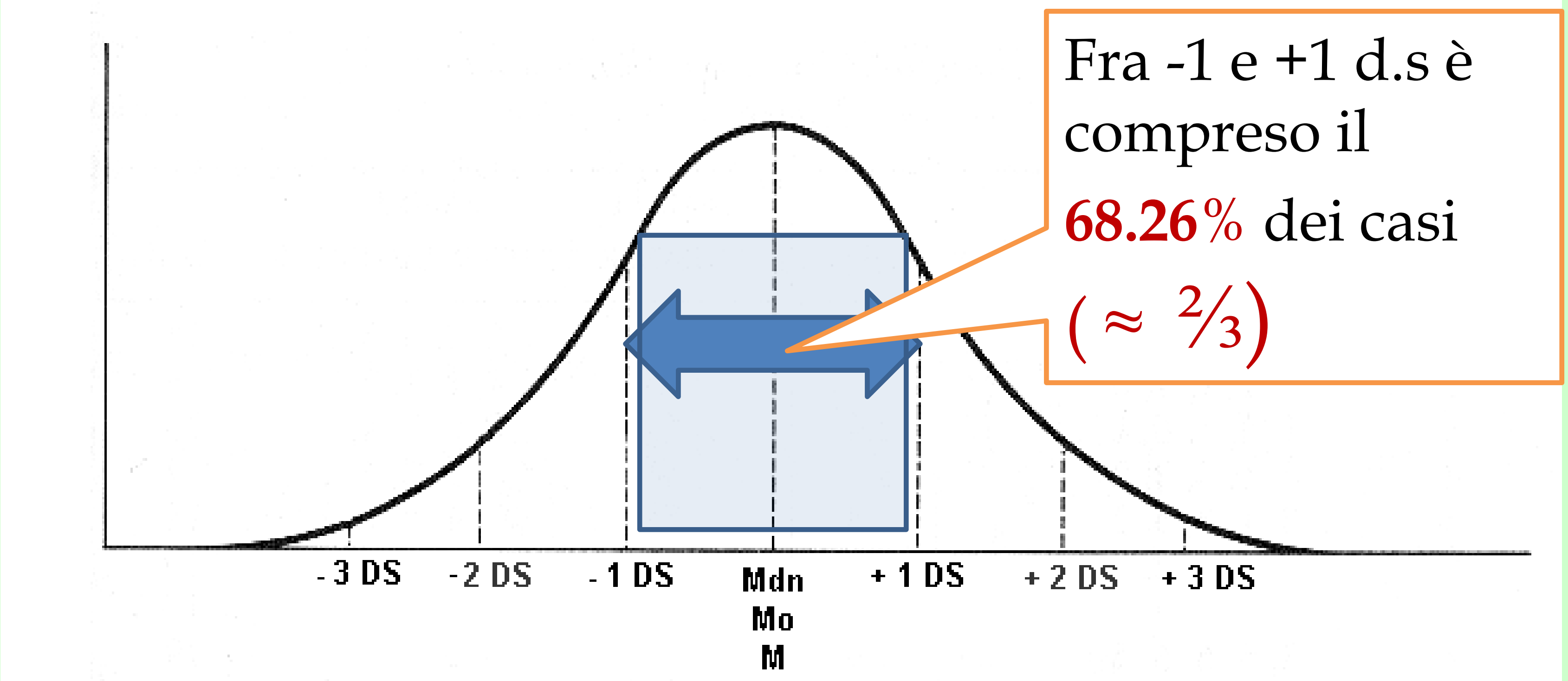


Sempre applicando ai punteggi di un test, la probabilità P che i soggetti abbiano realizzato un punteggio standardizzato x **maggiore** di un certo valore z si trova sottraendo a 0.5 la probabilità che il loro punteggio sia compreso tra 0 e z.



Dato che, come si è detto, la curva è simmetrica rispetto alla media, anche le aree sono simmetriche rispetto ad essa. L’area compresa tra un punto z negativo e la coda di sinistra sarà quindi equivalente all’area compresa tra lo stesso punto z (ma con segno positivo) e la coda di destra.

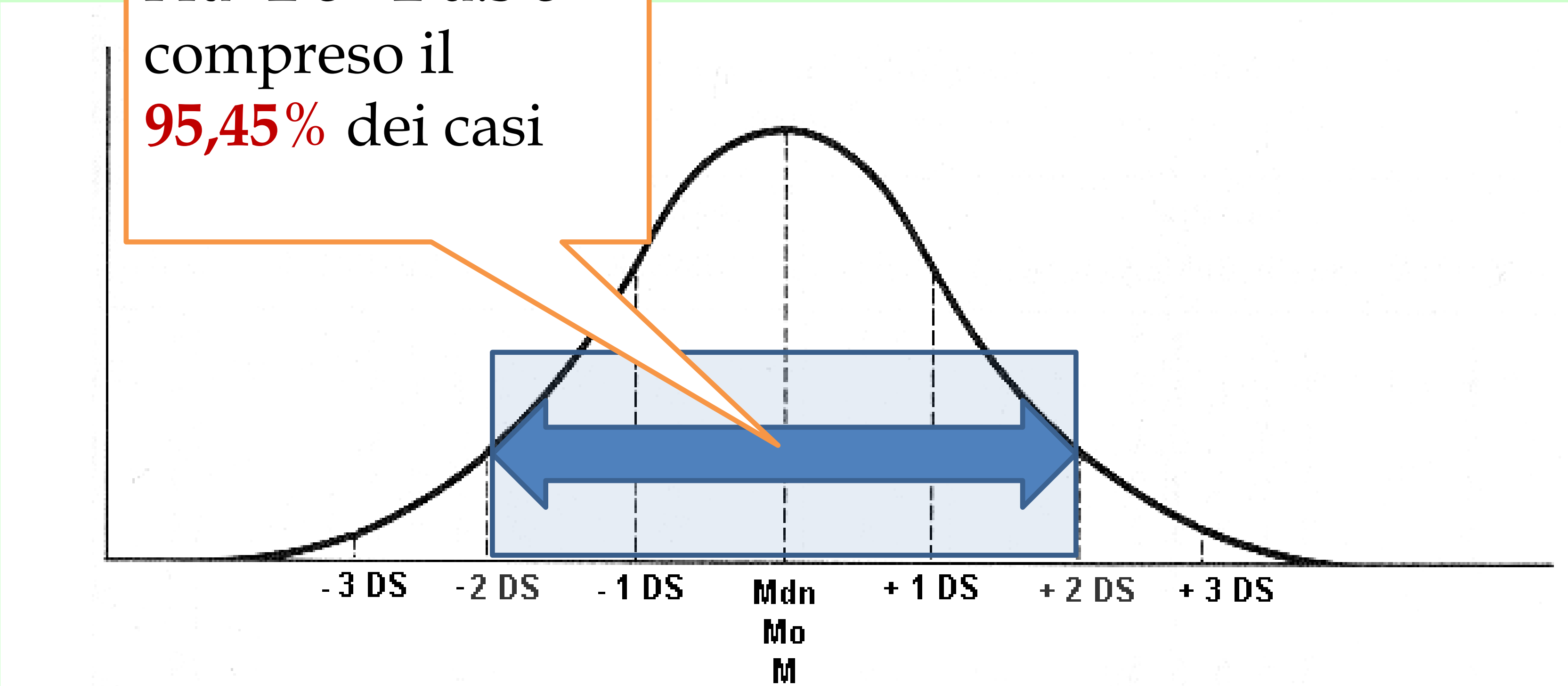
VALORI NOTEVOLI



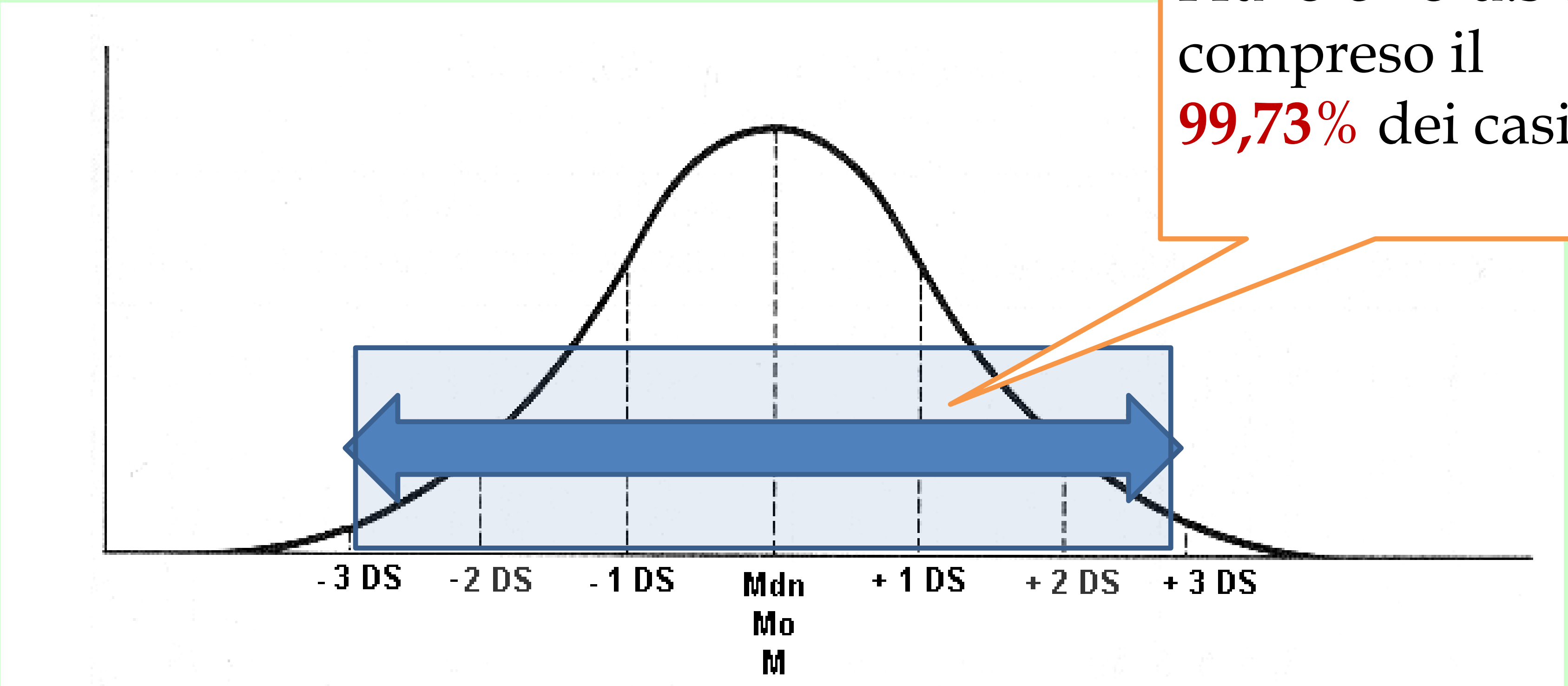
Fra -1 e +1 d.s è compreso il **68.26%** dei casi (≈ ⅔)

VALORI NOTEVOLI

Fra -2 e +2 d.s è compreso il **95,45%** dei casi



VALORI NOTEVOLI



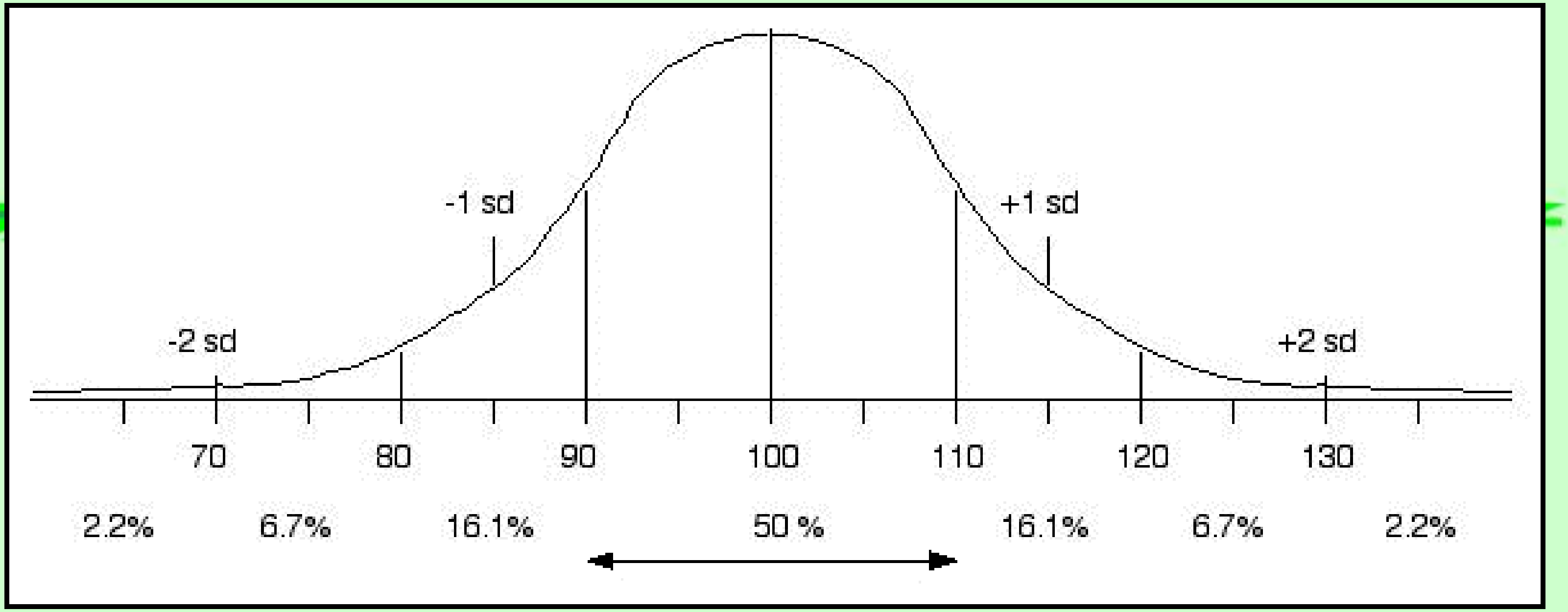
Fra -3 e +3 d.s è compreso il **99,73%** dei casi

Ancora…

Altri valori notevoli

- Il 90 % dei casi è compreso fra ± **1, 64** z
- Il 95 % dei casi è compreso fra ± **1, 96** z
- Il 50 % dei casi è compreso fra ± **0, 67** z

Questo valore ti ricorda qualcosa?



Under 70	**[mentally retarded] -- 2.2%**
70-80	**[borderline retarded] -- 6.7%**
80-90	**[low average] -- 16.1%**
90-110	**[average] -- 50%**
110-120	**[high average] -- 16.1%**
120-130	**[superior] -- 6.7%**
Over 130	**[very superior] -- 2.2%**

Esempio di calcolo applicativo

In un test Carlo ha avuto un punteggio pari a **23**. La media del test è **20** e la d.s. è uguale a **3**. Quante persone hanno ottenuto un punteggio migliore del suo?

SVOLGIMENTO

$X \sim N(\mu=20,\ \sigma=3)$

$P(X>23)=P(z>?)$

$z=(23-20)/3=3/3=1{,}00$

$P(X>23)=P(z>1{,}00)=0{,}5 - P(0<z<1{,}00)=0{,}5-0{,}3413=0{,}1587$

Percentuale = 15,87%

Risposta: il 15,87% delle persone ottiene un punteggio migliore di Carlo

ESERCIZIO N.2

Rossana ha avuto un punteggio di **114** in un test di abilità verbale che ha una media di 100 e una d.s. di 5. Quante persone (in percentuale) ottengono un punteggio più basso di Rossana? E di Alberto, che ha ottenuto un punteggio pari a **90**?

SVOLGIMENTO

$X \sim N(\mu=100, \sigma^2=25)$

Per Rossana

$P(X<114)=P(z<?)$ $z=(114-100)/5=14/5 = 2,8$

$P(X<114)=P(z<2,8)=0,5+P(0<z<2,8)=0,5+0,4974=0,9974$

Risposta: 99,74% ha ottenuto punteggi più bassi di Rossana

Per Alberto

$P(X<90)=P(z<?)$ $z=(90-100)/5= -10/5 = -2$

$P(X<90)=P(z<-2)=P(z>2)=0,5-P(0<z<2)=0,5-0,4772==0,0228$

Risposta: 2,28% ha ottenuto punteggi più bassi di Alberto

La curva gaussiana con Excel

- Il foglio di calcolo Excel (e altri simili) dispone, fra le sue funzioni, di una che permette di calcolare direttamente la percentuale (o meglio, quota) al di sopra di un dato punto z , oppure di un qualsiasi punto grezzo, di cui si conosce media e deviazione standard.
- Per conoscere le funzioni, che variano a seconda della versione di Excel, si può guardare fra le funzioni statistiche.

La curva gaussiana con Excel

- Funzioni di Excel incollabili nel foglio elettronico

- Dal punto zeta alla percentuale
- =DISTRIB.NORM.ST(cella)
- =DISTRIB.NORM(cella)

Verificare la dizione esatta che varia con la versione di Excel

- Dalla percentuale al punto zeta (o grezzo)
- =INV.NORM.ST(cella)
- =INV.NORM(cella)

Nella colonna B sono inseriti dei punti zeta, da -3 a 1,4, compreso il valore 0,52 dell'esempio precedente.

Nella colonna C è stata incollata la funzione

= DISTRIB.NORM.ST.N(cella; VERO)

che restituisce la quota di casi compresi dal punto più basso (infinito negativo) al punto zeta indicato nella cella di colonna B.

L'indicatore logico VERO dice che è VERO che si tratta di una funzione CUMULATIVA.

	A	B	C	D	E
1			=DISTRIB.NORM.ST.N(B4;VERO)		
2		-3	0,00135		
3		-2,8	0,00256		
4		-2,2	0,0139		
5		-2	0,02275		
6		-1,8	0,03593		
7		-1,2	0,11507		
8		-1	0,15866		
9		-0,6	0,27425		
10		-0,2	0,42074		
11		0	0,5		
12		0,4	0,65542		
13		0,8	0,78814		
14		1	0,84134		
15		1,2	0,88493		
16		1,4	0,91924		
17		0,52	0,69847		
18					

- Si può usare anche la funzione che accetta i valori dei punti grezzi: valore osservato, media, deviazione standard.

Rossana	=DISTRIB.NORM.N(114;100;5;VERO			
	0,9974	= 99,74%		
Alberto	=DISTRIB.NORM.N(90;100;5;VERO)			
	0,0228	=2,28 %		

- La quota può essere trasformata in percentuale

IN CASO DI DISTRIBUZIONE NON NORMALE

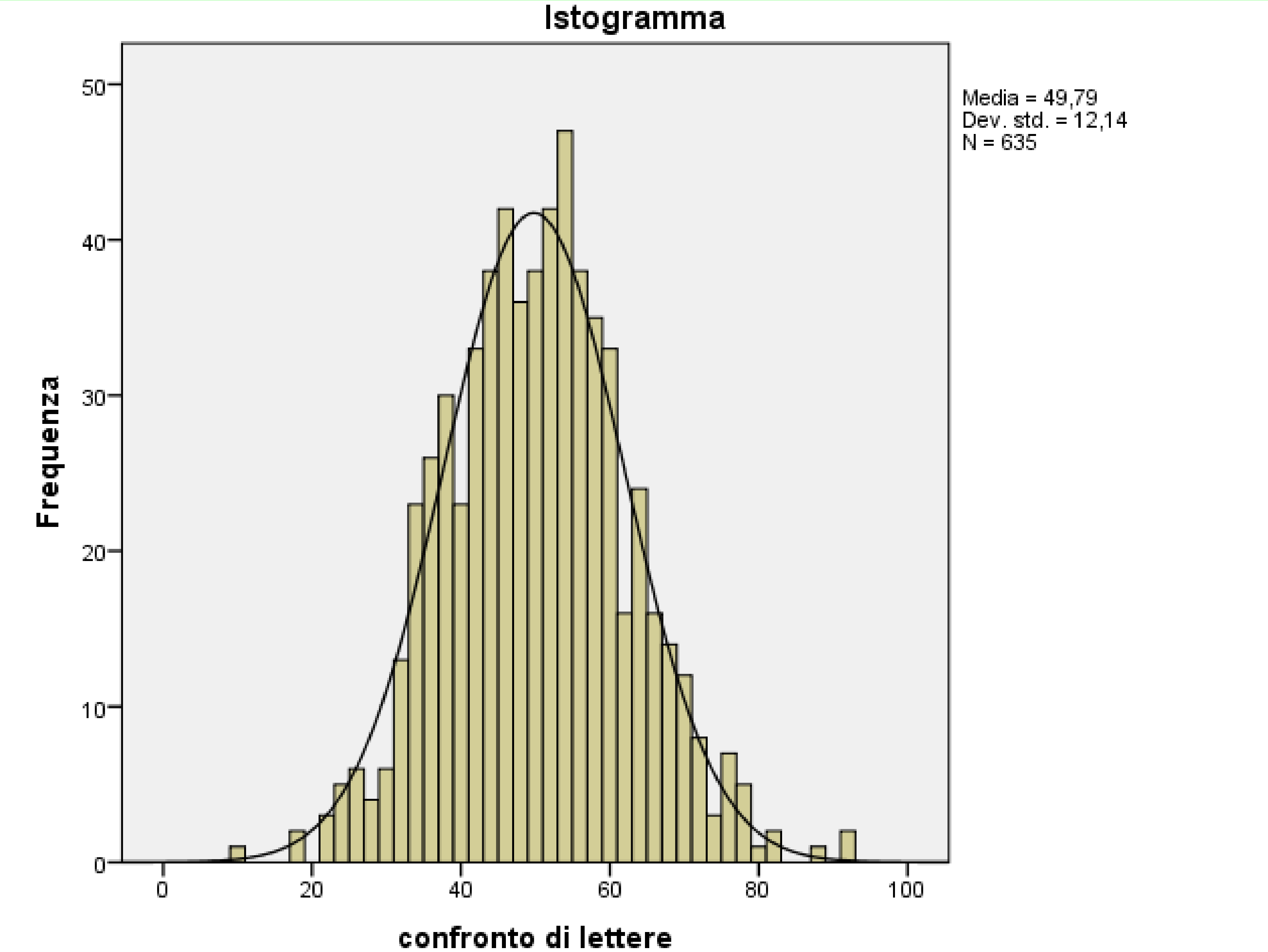
La maggioranza dei test e misurazioni usate in psicologia hanno delle distribuzioni approssimativamente normali.

Se la distribuzione non è normale, si può ricorrere alla perequazione (ridistribuzione dei punteggi secondo la normale) oppure usare i percentili.

Valori anomali

- E' possibile anche che la distribuzione sia accettabilmente normale, ma siano presenti dei valori estremi (molto bassi o molto alti) dovuti a diverse cause. Di solito è opportuno rimuoverli dalle analisi dei dati
- I valori anomali (English : outliers) si possono identificare con SPSS per la lontananza dalla media. SPSS fornisce alcuni strumenti per identificarli

Strumenti di SPSS per verificare la distribuzione delle osservazioni



Istogramma del test G1 con curva normale sovrapposta

SPSS produce i percentili

Percentili

		Percentili						
		5	10	25	50	75	90	95
Media pesata (definizione 1)	confronto di lettere	31,00	34,60	42,00	50,00	58,00	65,40	70,00
Cardini di Tukey	confronto di lettere			42,00	50,00	58,00		

SPSS stampa i casi estremi

Il file dei dati è stato prima riordinato secondo la variabile **Confronto di lettere-G1**
Ma non è necessario farlo

Valori estremi

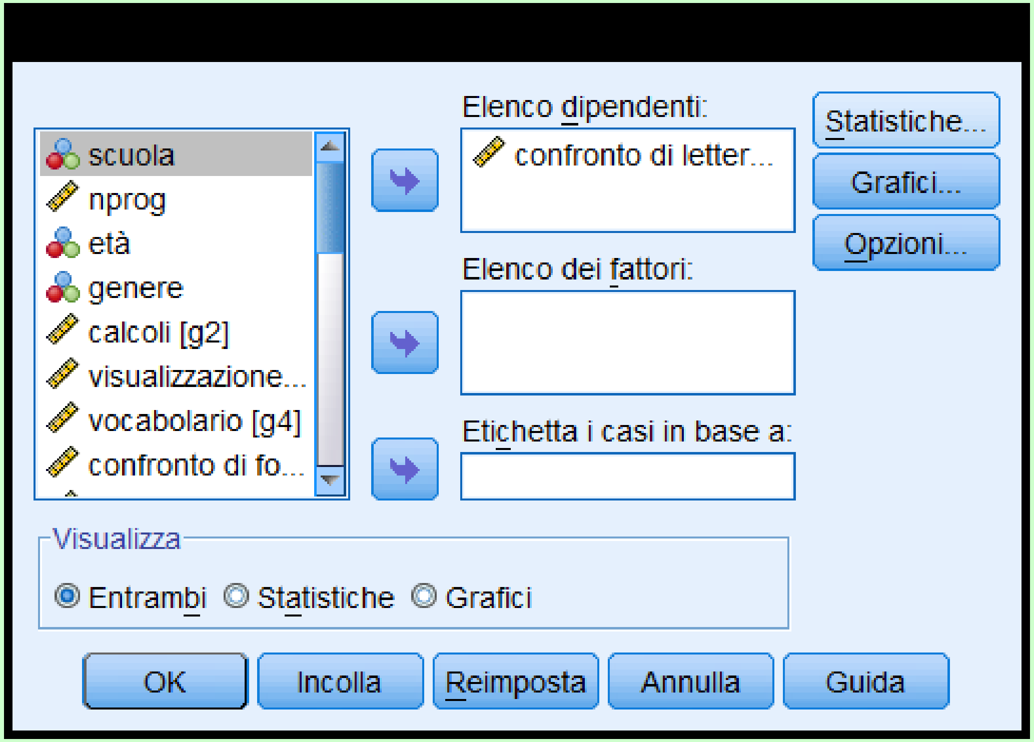
			Numero caso	Valore
confronto di lettere	Più alto	1	635	92
		2	634	91
		3	633	88
		4	631	82
		5	632	82
	Più basso	1	1	10
		2	3	17
		3	2	17
		4	4	21
		5	6	22[a]

a. Viene visualizzato solo un elenco parziale di casi con il valore 22 nella tabella degli estremi inferiori.

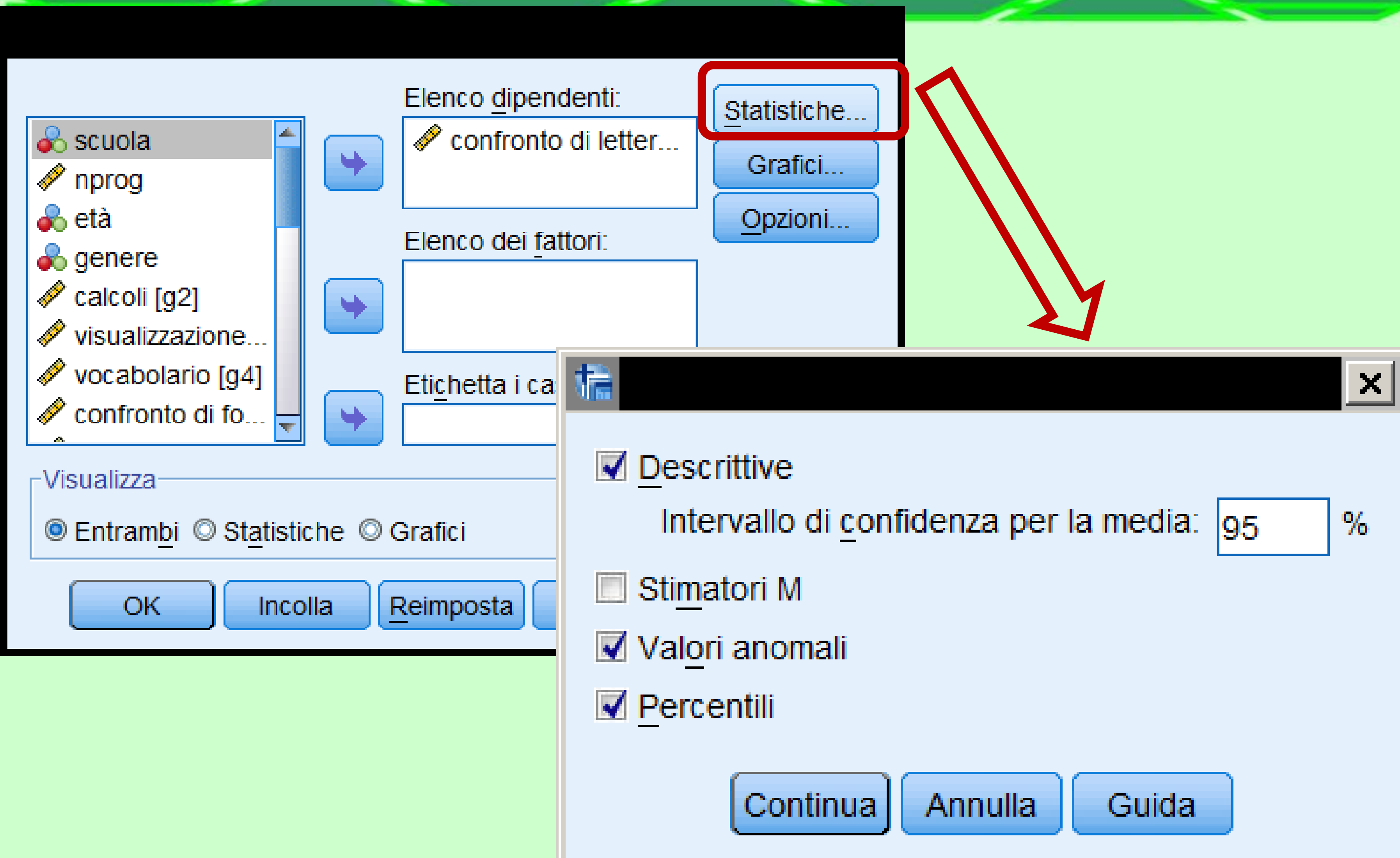
Grafico ramo-foglia

```
Confronto di lettere Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     3,00 Extremes    (=<17)
     8,00        2 .  244&
    14,00        2 .  566899&
    38,00        3 .  011112233333444444
    69,00        3 .  555555566666677777778888888999999
    81,00        4 .  00000111111112222222233333334444444444444
   103,00        4 .  55555555555566666666777777777777888889999999999999
   102,00        5 .  000000111111111112222222222333333333333333444444444
    88,00        5 .  55555555555666666667777777778888888899999999
    58,00        6 .  00000000011112223333333444
    37,00        6 .  555566667778888999
    16,00        7 .  001123
    13,00        7 .  55678&
     2,00        8 .  2
     3,00 Extremes    (>=88)

 Stem width:   10
 Each leaf:        2 case(s)


 & denotes fractional leaves.
```

Tabelle e grafici si ottengono dal menù **Analizza**
Poi **Esplora** (English:Examine)
e infine i due pulsanti **Statistiche** e **Grafici**



Menu Analizza



Test di normalità						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistica	gl	Sign.	Statistica	gl	Sign.
confronto di lettere	,032	635	$\boxed{,184}$	,997	635	$\boxed{,303}$

a. Correzione di significatività di Lilliefors

I due test di normalità, (Kolmogorov Smirnov e Shapiro Wilk) non sono significativi e permettono di concludere che non vi è una deviazione apprezzabile dalla normalità. Pertanto la distribuzione del test G1 può essere considerata normale

La perequazione (smoothing)

- Quando i dati non hanno una distribuzione accettabilmente normale, si può aggiustarli e trasformarli leggermente.
- La perequazione (smoothing in inglese) verrà presentata successivamente