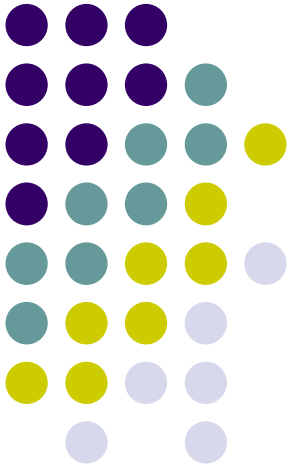
Levine, Krehbiel, Berenson

Statistica

Capitolo 6

La distribuzione normale



Insegnamento: Statistica (gruppo C)

Corso di Laurea Triennale in Economia

Università degli Studi di Ferrara

Docente: Dott.ssa A. Grassi

Si ringrazia il Prof. S. Bonnini per aver condiviso le slide del suo corso

Argomenti

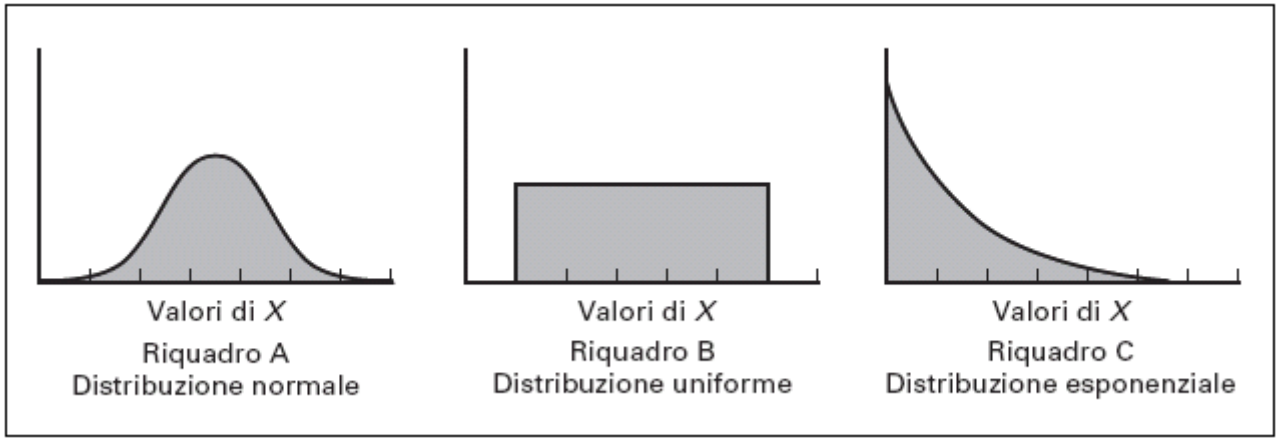
- Distribuzioni di probabilità continue
- La distribuzione normale
  - importanza e caratteristiche della distribuzione normale
  - la standardizzazione
  - ricavare dalle tavole le probabilità desiderate
  - determinare il valore associato a una data probabilità
- Valutazione dell’ipotesi di normalità

Distribuzioni di probabilità continue

- Una **funzione di densità di probabilità continua** è un modello che definisce analiticamente come si distribuiscono i valori assunti da una variabile aleatoria continua.
- Quando si dispone di un'espressione matematica adatta alla rappresentazione di un fenomeno continuo, siamo in grado di calcolare la probabilità che la variabile aleatoria assuma valori compresi in intervalli.
- Tuttavia la probabilità che la variabile aleatoria continua assuma *un particolare valore* è pari a zero.
- I modelli continui hanno importanti applicazioni in ingegneria, fisica, economia e nelle scienze sociali.

Distribuzioni di probabilità continue

- Alcuni tipici fenomeni continui sono l'altezza, il peso, le variazioni giornaliere nei prezzi di chiusura di un'azione, il tempo che intercorre fra gli arrivi di aerei presso un aeroporto, il tempo necessario per servire un cliente in un negozio.
- La figura rappresenta graficamente tre funzioni di densità di probabilità: normale, uniforme ed esponenziale.



La distribuzione normale

La distribuzione **normale** (o distribuzione *Gaussiana*) è la distribuzione continua più utilizzata in statistica.

La distribuzione normale è importante in statistica per tre motivi fondamentali:

1. Diversi fenomeni continui sembrano seguire, almeno approssimativamente, una distribuzione normale.
2. La distribuzione normale può essere utilizzata per approssimare numerose distribuzioni di probabilità discrete.
3. La distribuzione normale è alla base dell'*inferenza statistica classica* in virtù del *teorema del limite centrale* (paragrafo 7.2).

La distribuzione normale

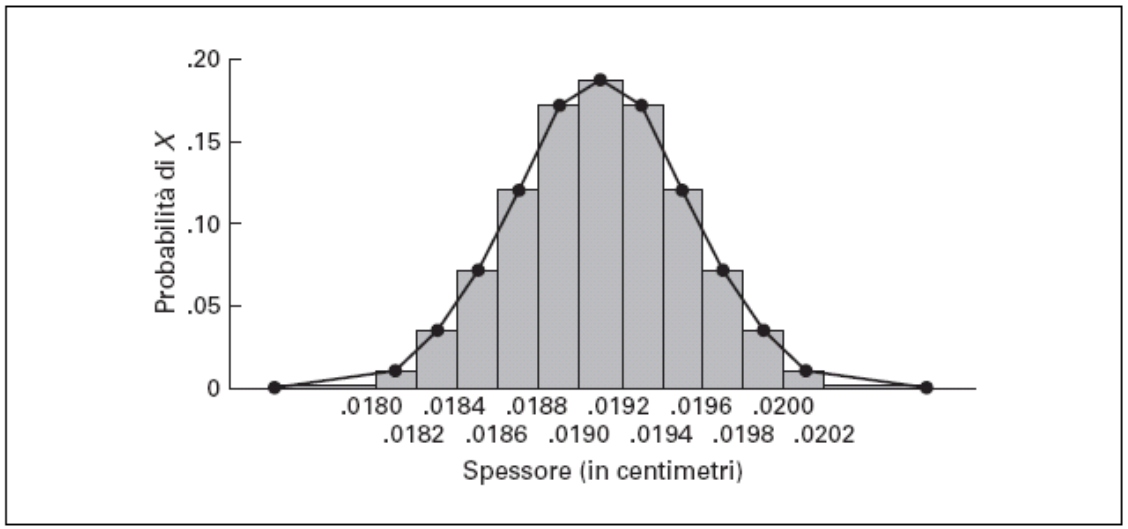
La distribuzione normale ha alcune importanti caratteristiche:

- La distribuzione normale ha una forma campanulare e simmetrica.
- Le sue misure di posizione centrale (valore atteso, mediana) coincidono.
- Il suo range interquartile è pari a 1.33 volte lo scarto quadratico medio, cioè copre un intervallo compreso tra $\mu - 2/3\sigma$ e $\mu + 2/3\sigma$.
- La variabile aleatoria con distribuzione normale assume valori compresi tra $-\infty$ e $+\infty$.

La distribuzione normale

Molte variabili statistiche che osserviamo nella realtà hanno una distribuzione con caratteristiche simili a quelle della distribuzione normale.

Consideriamo ad esempio lo spessore misurato in centimetri di 10000 rondelle di ottone prodotte da una grande società metallurgica. Il fenomeno aleatorio continuo di interesse, lo spessore delle rondelle, si distribuisce approssimativamente come una normale.



La distribuzione normale

Utilizzeremo il simbolo *f*(*X*) per denotare l'espressione matematica di una funzione di densità di probabilità. Nel caso della distribuzione normale la **funzione di densità di probabilità normale** è data dalla seguente espressione:

Funzione di densità di probabilità normale

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2)[(X-\mu)/\sigma]^2} \qquad \textbf{(6.1)}$$

dove *e* = costante matematica approssimata da 2.71828

π = costante matematica approssimata da 3.14159

μ = valore atteso della popolazione

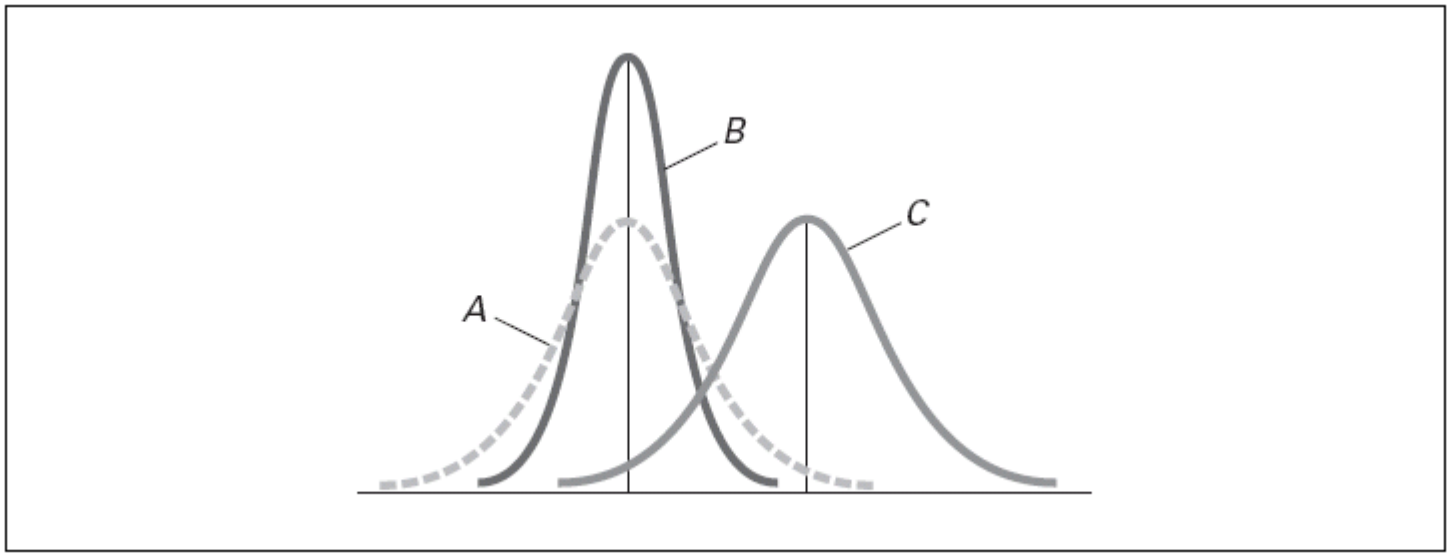
σ = scarto quadratico medio della popolazione

X = valori assunti dalla variabile aleatoria, $-\infty < X < +\infty$

La distribuzione normale

Notiamo che, essendo *e* e π delle costanti matematiche, le probabilità di una distribuzione normale dipendono soltanto dai valori assunti dai due parametri μ e σ.

Specificando particolari combinazioni di μ e σ, otteniamo differenti distribuzioni di probabilità normali.



La distribuzione normale

Poiché esiste un numero infinito di combinazioni dei parametri μ e σ, per poter rispondere a quesiti relativi a una qualsiasi distribuzione normale avremmo bisogno di in numero infinito di tavole.

Introduciamo ora una formula di trasformazione delle osservazioni, chiamata standardizzazione, che consente appunto di trasformare una generica variabile aleatoria normale in una variabile aleatoria normale standardizzata.

La standardizzazione

$$Z = \frac{X - \mu}{\sigma} \qquad \textbf{(6.2)}$$

Z è la variabile ottenuta sottraendo ad *X* il suo valore atteso μ e rapportando il risultato allo scarto quadratico medio, σ.

La distribuzione normale

La variabile aleatoria standardizzata *Z* ha la caratteristica di avere valore atteso nullo (μ=0) e scarto quadratico medio pari a uno (σ=1).

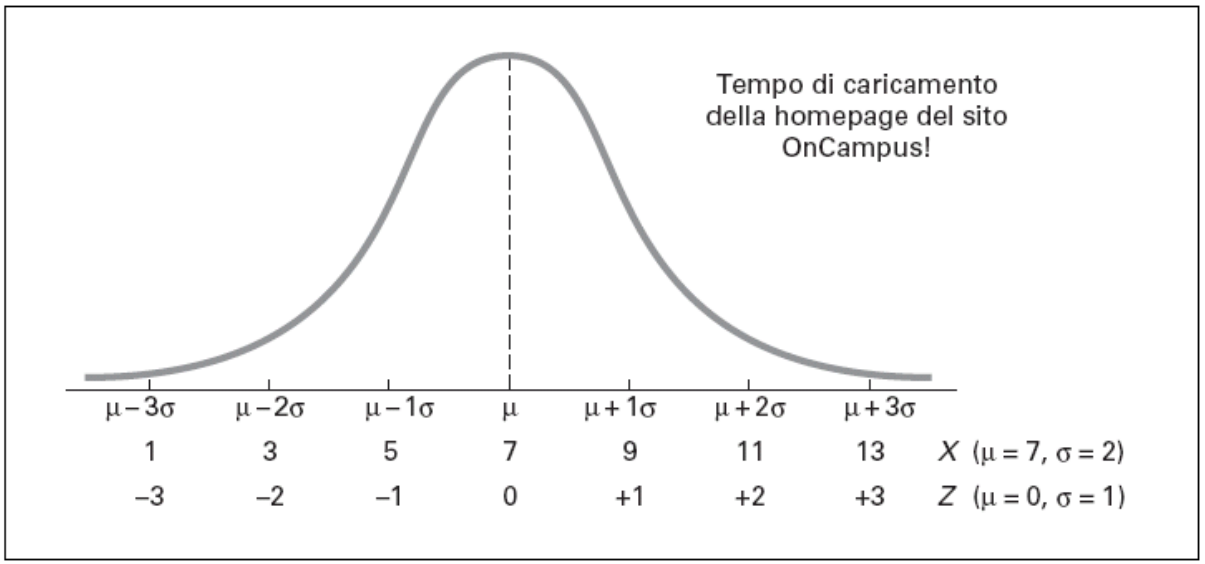
Quindi è sempre possibile trasformare qualsiasi insieme di valori distribuiti normalmente nel corrispondente insieme di valori standardizzati e ricavare le probabilità desiderate dalle tavole della distribuzione normale standardizzata (Tavole E.2(a) e E.2(b)).

Supponiamo che il tempo necessario per caricare la home page del sitoOnCampus! sia distribuito normalmente con μ=7 secondi e scarto quadratico medio pari σ=2 secondi.

La distribuzione normale

Nella figura si osserva come a ciascun valore della variabile *X* (tempo di caricamento) è associato il corrispondente valore della variabile standardizzata *Z*, ottenuto applicando l'equazione (6.2).

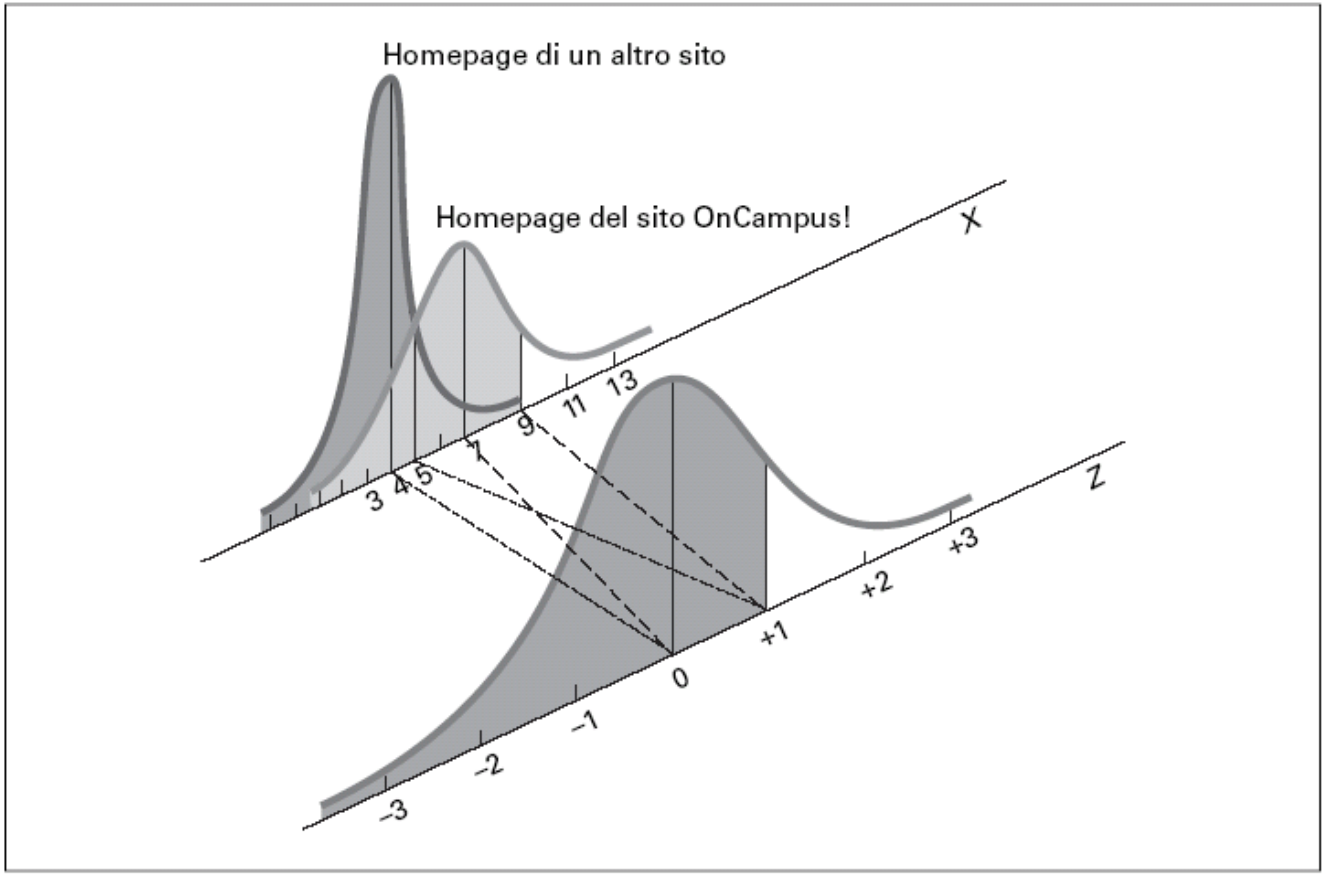
Supponiamo di voler determinare la probabilità che il tempo di caricamento della home page in una generica sessione sia inferiore ai 9 secondi.



La distribuzione normale

Applicando l'equazione (6.2). si ottiene che a X=9 corrisponde il valore della variabile standardizzata

$$Z=\frac{(9-7)}{2}=+1$$

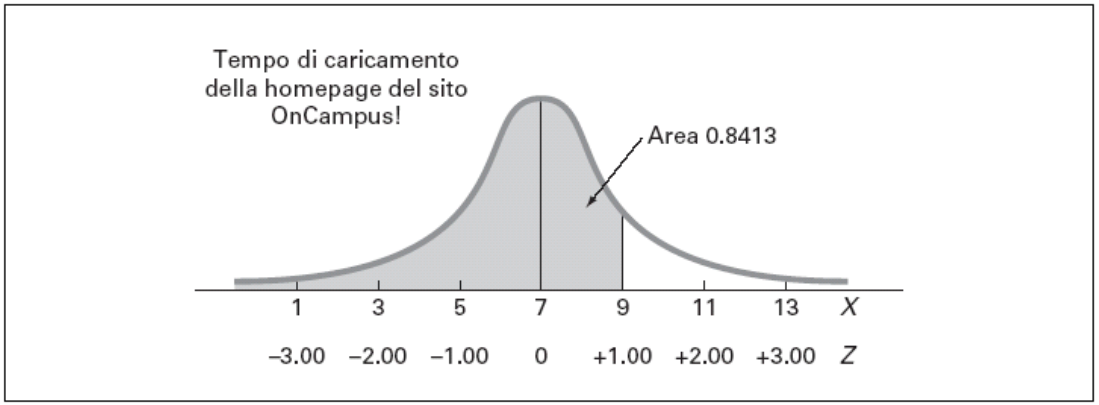


La distribuzione normale

Dopodiché si utilizza la Tavola E.2 per determinare l'area cumulata fino al valore 1.

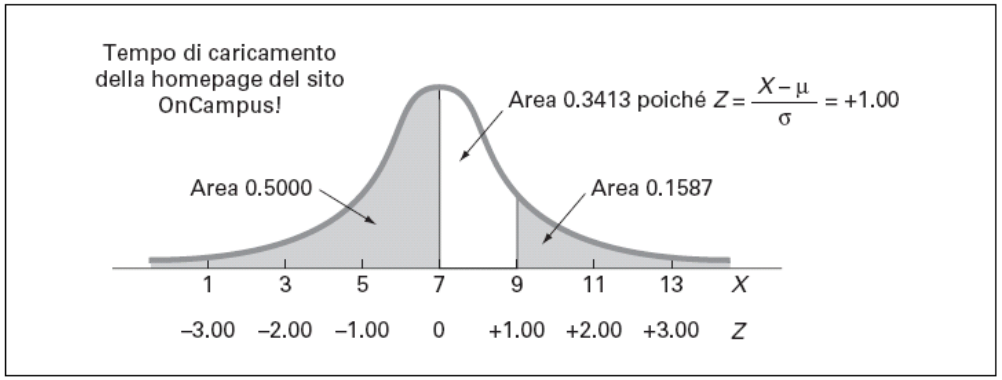
Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7518	.7549
0.7	.7580	.7612	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0 →	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621

Fonte: estratto dalla Tavola E.2

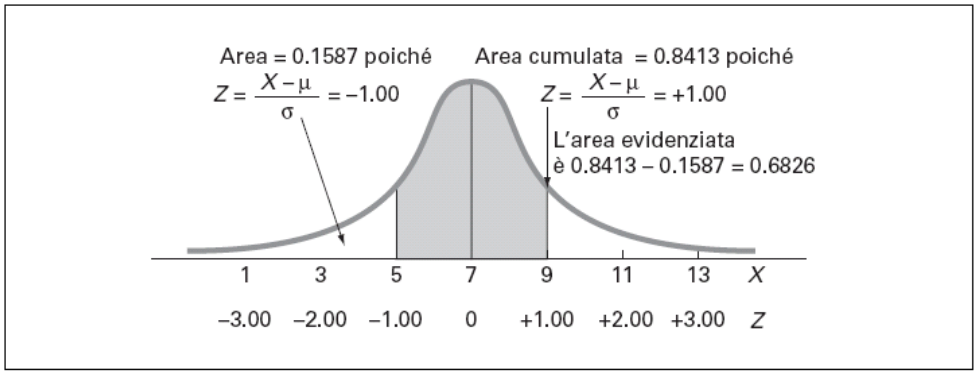


La distribuzione normale

Esempio 6.1 Tempo di caricamento della home page del sito OnCampus!: calcolo di $P(X<7 \text{ o } X>9)$

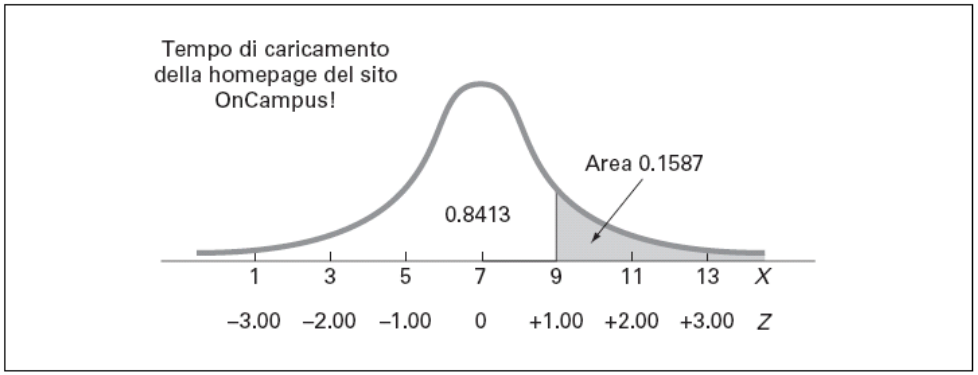


Esempio 6.2 Tempo di caricamento della home page del sito OnCampus!: calcolo di $P(5<X<9)$

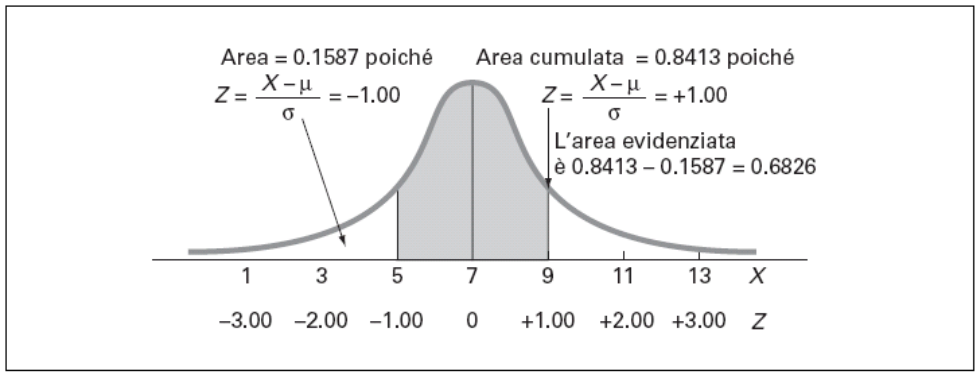


La distribuzione normale

Esempio 6.3 Tempo di caricamento della home page del sito OnCampus!: calcolo di $P(X>9)$



Esempio 6.4 Tempo di caricamento della home page del sito OnCampus!: calcolo di $P(5<X<9)$



La distribuzione normale

Il risultato dell'esempio 6.4 può essere generalizzato, infatti per un insieme di dati con distribuzione normale:

- approssimativamente il 68.26% apparterrà all'intervallo $(\mu - \sigma, \mu + \sigma)$
- approssimativamente il 95.44% apparterrà all'intervallo $(\mu - 2\sigma, \mu + 2\sigma)$
- approssimativamente il 99.73% apparterrà all'intervallo $(\mu - 3\sigma, \mu + 3\sigma)$

È quindi evidente il motivo per cui un intervallo di ampiezza 6σ centrato su μ, vale a dire l'intervallo $(\mu - 3\sigma, \mu + 3\sigma)$, può essere considerato come un'*approssimazione pratica del range* per dati distribuiti normalmente.

La distribuzione normale

Negli esempi 6.1-6.4 la tavola della distribuzione normale standardizzata viene utilizzata per calcolare l'area fino ad un certo valore *X*. In molte applicazioni si è però interessati al procedimento opposto, cioè determinare il valore di *X* cui corrisponde una certa area cumulata.

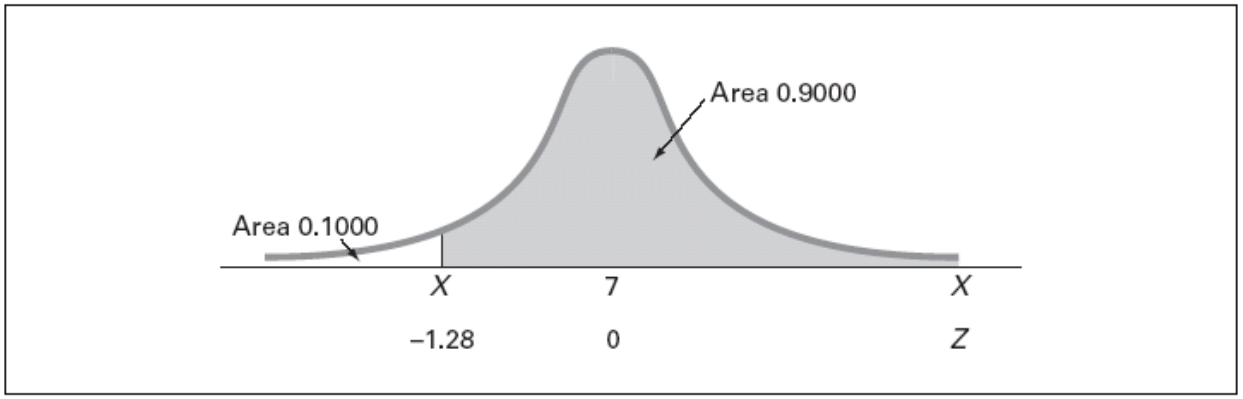
Esempio 6.6 Tempo di caricamento della home page del sito OnCampus!: calcolo del tempo massimo di caricamento per almeno il 10% delle sessioni

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.0156	.0138	.1020	.1003	.0985

Fonte: estratto dalla Tavola E.2

La distribuzione normale

Esempio 6.6 Tempo di caricamento della home page del sito OnCampus!: calcolo del tempo massimo di caricamento di almeno il 10% delle sessioni



Determinare il valore X associato a una probabilità (cumulata)

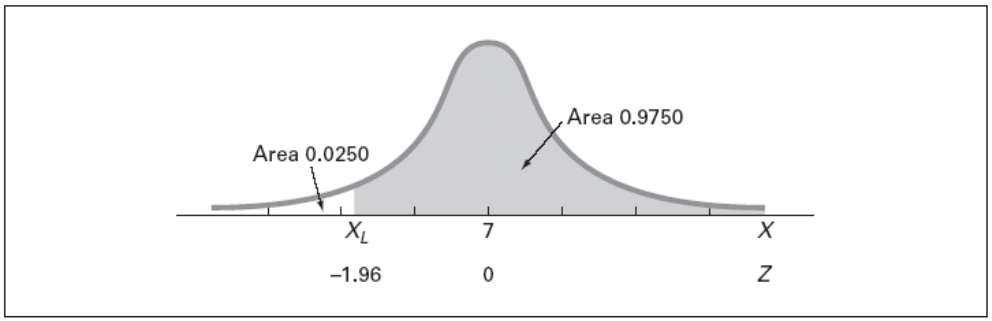
$$X = \mu + Z\sigma \qquad \textbf{(6.4)}$$

il valore *X* è dato dalla media μ, cui va sommato il prodotto tra *Z* e lo scarto quadratico medio, σ.

$X = 7 + (-1.28)(2) = 4.44$ secondi

La distribuzione normale

Esempio 6.7 Tempo di caricamento della home page del sito OnCampus!: determinazione dell'intervallo centrato sulla media al quale appartiene il 95% dei tempi di caricamento

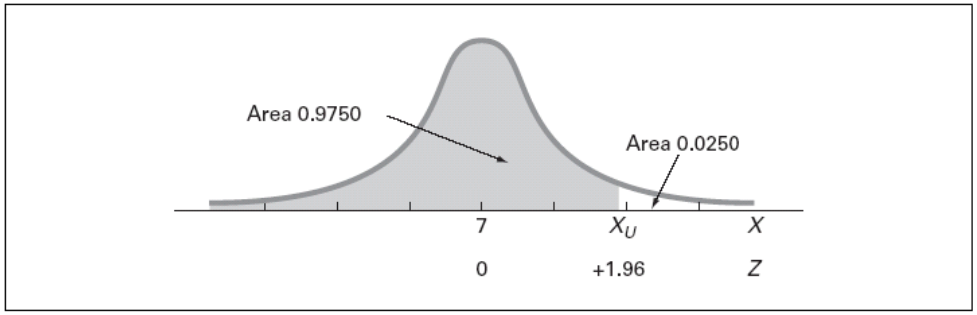


Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0232	.0314	.0307	.0301	.0294

Fonte: estratto dalla Tavola E.2

La distribuzione normale

Esempio 6.7 Tempo di caricamento della home page del sito OnCampus!: determinazione dell'intervallo centrato sulla media al quale appartiene il 95% dei tempi di caricamento



Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
+1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
+1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
+2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817

Fonte: estratto dalla Tavola E.2

$$X = 7 + (-1.96)(2) = 3.08 \text{ secondi}$$

$$X = 7 + (+1.96)(2) = 10.92 \text{ secondi}$$

Valutazione dell'ipotesi di normalità

Non tutti i fenomeni continui sono distribuiti normalmente e non tutti seguono una distribuzione che può essere approssimata adeguatamente con una normale. È quindi importante verificare la plausibilità dell'ipotesi di normalità, cioè di accertare se in effetti un insieme di dati può provenire da una distribuzione normale. Dal punto di vista pratico il problema è di valutare la bontà di adattamento del modello normale a un insieme di dati, problema che deve essere affrontato ancora prima di applicare le metodologie descritte nel precedente paragrafo.

Due sono gli approcci esplorativi di carattere descrittivo che possono essere adottati:

1. Il confronto fra le caratteristiche dei dati e le proprietà di un'eventuale distribuzione normale sottostante.
2. La costruzione di un normal probability plot.

Valutazione dell'ipotesi di normalità

La distribuzione normale ha alcune importanti proprietà teoriche:

- è simmetrica: la media e la mediana coincidono;
- ha forma campanulare, di modo che può essere applicata la regola empirica;
- il suo range interquartile è pari a 1.33 volte lo scarto quadratico medio;
- il range è infinito.

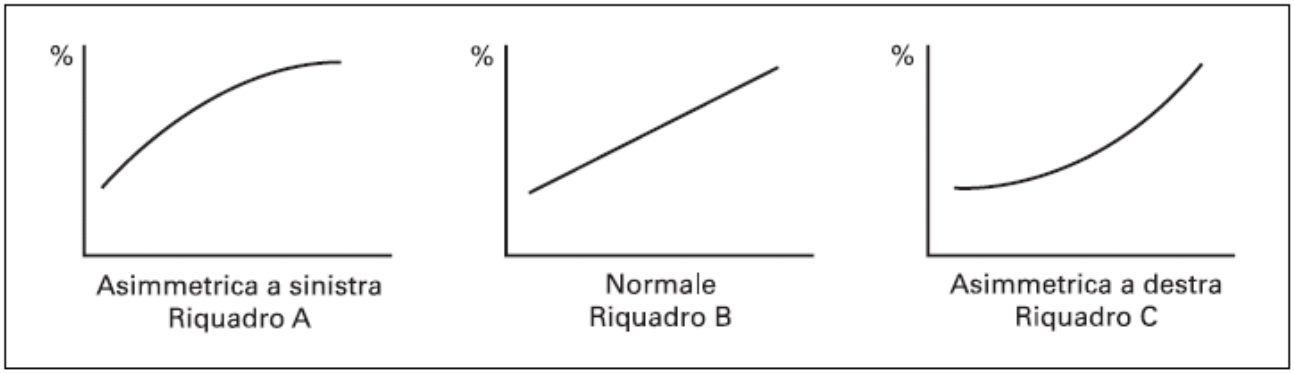
Per un dato insieme di dati, per valutare l'adeguatezza dell'ipotesi di normalità si può procedere con

- la costruzione di grafici per analizzare la forma della distribuzione;
- il calcolo delle misure di sintesi e il confronto con le proprietà teoriche;
- il confronto fra le caratteristiche dei dati e le proprietà di un'eventuale distribuzione normale sottostante.

Valutazione dell'ipotesi di normalità

Un **normal probability plot** è un grafico a due dimensioni in cui le osservazioni sono riportate sull'asse verticale e a ciascuna di esse viene fatto corrispondere sull'asse orizzontale il relativo quantile di una distribuzione normale standardizzata.

Se i punti del grafico si trovano approssimativamente su una linea retta immaginaria inclinata positivamente, allora possiamo affermare che i dati osservati si distribuiscono approssimativamente secondo la legge normale.



Valutazione dell'ipotesi di normalità

Figura 6.22
Normal Probability Plot per il rendimento 2003 dei fondi comuni di investimento ottenuto con Microsoft Excel