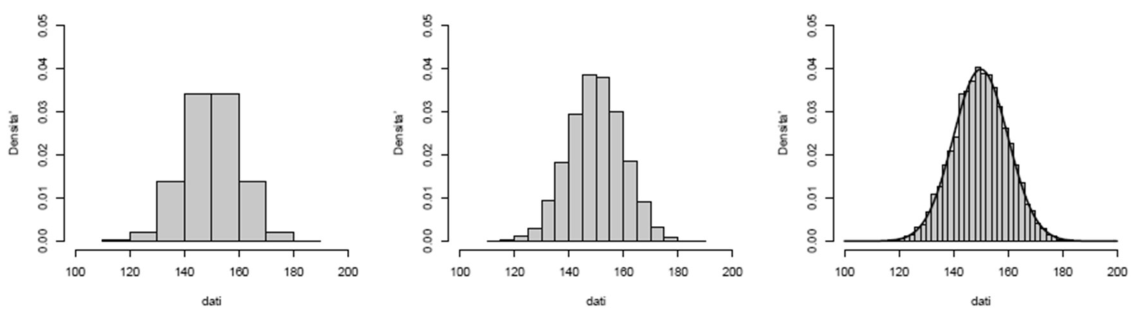
Curve di densità

La funzione di densità può essere considerata come il limite a cui tende un istogramma discreto.

Esempio

Sia X la misura dell'altezza delle persone.

- Possiamo rendere discreta X misurando l'altezza in centimetri, oppure in millimetri.
- Tanto più precise saranno le nostre misurazioni, tanto più l'istogramma che potremmo produrre avendo a disposizione un grande numero di misure si approssimerà ad una curva continua.
- Questa curva continua è la funzione di densità di probabilità (fdp).



- Questa curva continua si chiama funzione di densità e ha le seguenti proprietà:
  - l'area totale sotto la curva di densità è uguale a 1.0;
  - la proporzione di dati della distribuzione con un valore compreso in un certo intervallo è uguale all'area sottesa alla curva di densità in quell'intervallo;
  - la curva di densità non assume mai valori negativi (...sull'ordinata, naturalmente).

- Per trovare l'area sottesa alla funzione di densità sono necessari calcoli matematici complessi (integrali).
- Per i nostri scopi, tuttavia, dobbiamo soltanto capire l'idea di base:

$$area \quad \rightarrow \quad proporzione\ di\ casi$$

La funzione cumulativa di densità $F(x)$ di una variabile aleatoria continua X per ogni numero x è definita da

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} p(x)\,\mathrm{d}x$$

Per ciascun x, $F(x)$ è l'area sottesa alla curva alla sinistra di x.

Allora, per ogni numero a,

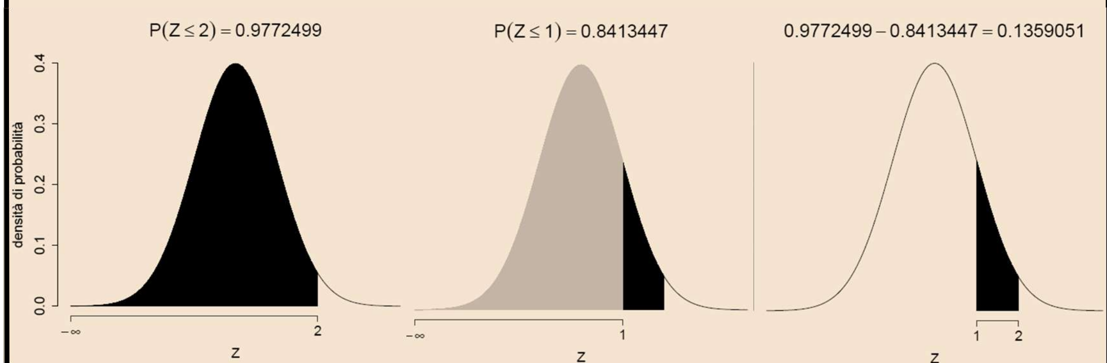
$$P(X > a) = 1 - F(a)$$

e per ogni coppia di numeri a e b, con $a < b$,

$$P(a \leq X \leq b) = F(b) - F(a)$$

Nella figura seguente è rappresentata l'area sottesa alla funzione densità di probabilità *normale standardizzata* nell'intervallo compreso tra 1 e 2. Tale area corrisponde alla probabilità $P(1 \leq Z \leq 2)$.

$$P(1 \leq Z \leq 2) = F(2.0) - F(1.0) =$$
$$0.9772 - 0.8413 = 0.1359 \text{ per } Z \sim N(0;1).$$

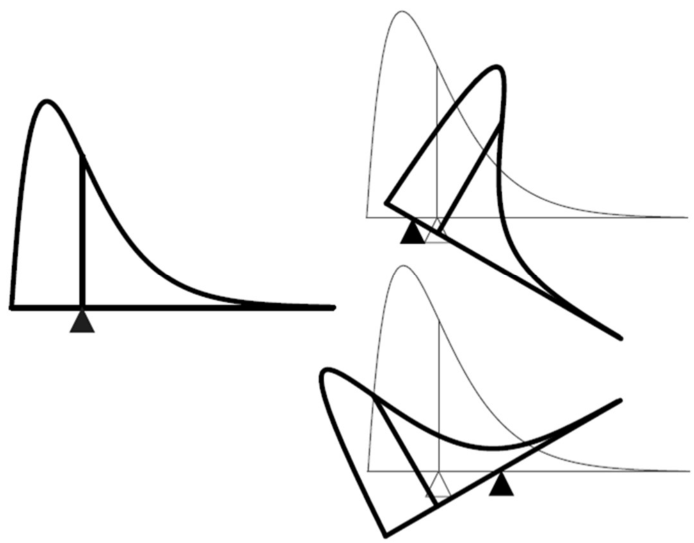


- Abbiamo imparato a calcolare la media, mediana e deviazione standard di un insieme di dati. Queste concetti hanno lo stesso significato nel caso di una funzione di densità.
  - La **media** di una curva di densità è il valore della variabile tale per cui metà dell'area sottesa alla curva si trova al di sotto di quel valore e metà si trova al di sopra di esso.
  - Se la curva di densità venisse ritagliata in un materiale solido (per esempio, un foglio di cartone), allora la media sarebbe il punto d'equilibrio.

notazione La media di una curva di densità è denotata da μ.

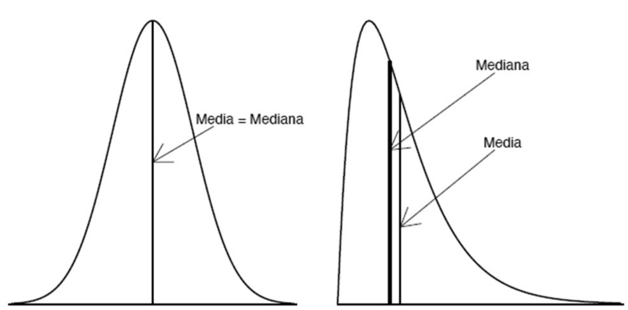
Le lettere greche sono tradizionalmente usate per indicare le caratteristiche delle funzioni di densità, per aiutarci a distinguere la media $\bar{x}$ di un campione di osservazioni dalla media μ di una funzione di densità.
Lo stesso dicasi per la varianza campionaria s^2 e la varianza della popolazione σ^2.

$\mu =$ punto di equilibrio



Media e Mediana

- In una distribuzione simmetrica la media e la mediana coincidono.
- In una distribuzione asimmetrica sia la media che la mediana si spostano nella direzione della coda più lunga della distribuzione —la media più della mediana.



Deviazione standard

- È possibile definire la deviazione standard, denotata dalla lettera σ anche per una curva di densità.
- La deviazione standard di una funzione di densità è più difficile da visualizzare della media ma, anche in questo caso, possiamo pensare alla deviazione standard di una serie di osservazioni come ad una sorta di scarto medio delle osservazioni dalla media della distribuzione.

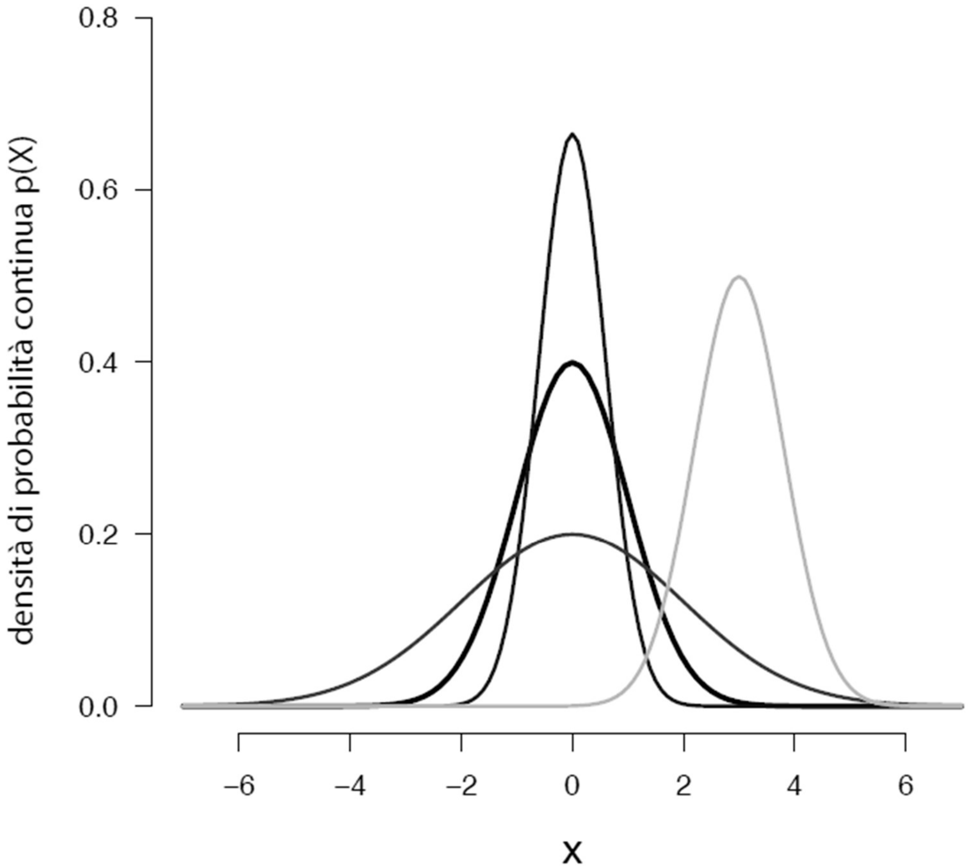
Le distribuzioni normali

- Le cosiddette distribuzioni normali sono le distribuzioni più importanti in statistica. Le distribuzioni normali sono anche dette gaussiane in onore di Carl Friedrich Gauss, uno dei matematici che le ha scoperte.
- Questo è un nome migliore perché, come le persone "normali", anche le variabili distribuite normalmente sono difficili da trovare.
- Perché allora le distribuzioni normali sono importanti per la statistica? Sono importanti per il ruolo centrale che hanno nell'inferenza statistica.

Micceri, T. (1989) *The unicorn, the normal curve, and other improbable creatures*. Psychological Bulletin, 105(1), 156-166.

- Tutte le distribuzioni gaussiane hanno la stessa forma: sono simmetriche, unimodali e a forma di campana.
- Le distribuzioni normali hanno medie e deviazioni standard diverse.
- C'è una diversa distribuzione normale per ciascuna coppia μ e σ.
- Una distribuzione normale con media μ e deviazione standard σ viene denotata da $N(\mu, \sigma)$

Le distribuzioni normali



Definizione

- In termini formali, diciamo che una variabile aleatoria continua ha una distribuzione normale con parametri μ e σ, dove $-\infty < \mu < \infty$ e $0 < \sigma$, se la funzione di densità di X è

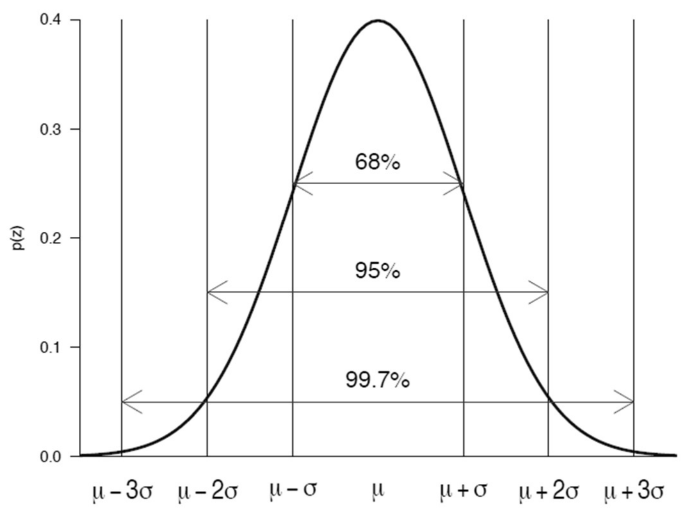
$$f(x;\mu;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}}; \quad (\vee x, -\infty < x < \infty)$$

- Può essere dimostrato che $E(X) = \mu$ e $V(X) = \sigma^2$.

Proprietà delle distribuzioni normali

Tutte le distribuzioni normali hanno le seguenti proprietà:

- Il 68% (circa due terzi) delle osservazioni è compreso nell'intervallo tra ± 1 deviazione standard σ dalla media μ.
- Il 95% delle osservazioni è compreso nell'intervallo tra ± 2 deviazioni standard σ dalla media μ.
- Il 99.7% delle osservazioni è compreso nell'intervallo tra ± 3 deviazioni standard σ dalla media μ



Illustrazione

Si considerino i punteggi d'esame di un grande numero di studenti. Si supponga che il punteggio medio sia 70 (in centesimi), con una deviazione standard di 5. Si supponga, inoltre, che i punteggi siano distribuiti in maniera approssimativamente normale.
Che percentuale di studenti riceve un punteggio compreso tra 60 e 80?

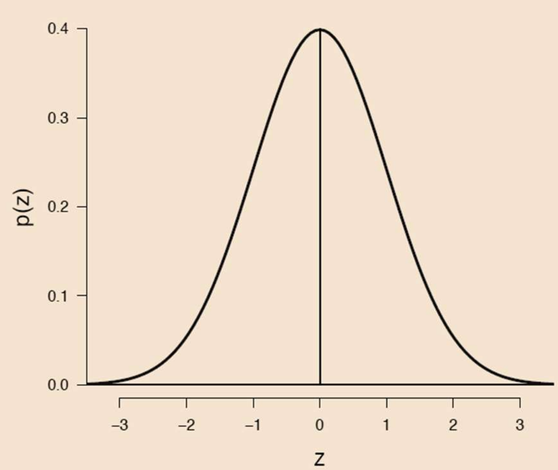
- Dato che 60 è 2 deviazioni standard sotto la media e 80 è 2 deviazioni standard sopra la media, circa il 95% degli studenti ha un punteggio compreso in questo intervallo.
- Circa il 2.5% degli studenti ha punteggi inferiori a 60 e circa il 2.5% ha punteggi superiori a 80.

La distribuzione normale standardizzata

- Tutte le distribuzioni normali si riducono alla stessa distribuzione se vengono misurate in unità σ attorno alla media μ. Tale mutamento dell'unità di misura della distribuzione viene detto standardizzazione.
- Anche se la nozione di standardizzazione è stata introdotta nel contesto delle distribuzioni normali, una variabile può essere standardizzata che se non è distribuita normalmente.

- Se x è un'osservazione facente parte di una distribuzione avente media μ e deviazione standard σ, allora il valore standardizzato di x si calcola come:
$$\frac{x-\mu}{\sigma}$$
- Le variabili standardizzate, dette punti z, hanno media 0 e varianza 1.

La distribuzione normale standardizzata si indica con $N(0,1)$. Una variabile aleatoria normale standardizzata viene denotata da Z.



- Nel contesto della distribuzione normale, la standardizzazione è utile in quanto trasforma qualsiasi distribuzione normale – con media μ e deviazione standard σ – in una distribuzione normale con media 0 e deviazione standard 1.
- Se la distribuzione di x è $N(\mu,\sigma)$, allora la distribuzione di $z=(x-\mu)/\sigma$ è $N(0,1)$.

Illustrazione

si considerino nuovamente i punteggi d'esame di un gruppo di studenti e supponiamo che il punteggio medio sia $\mu = 70$ con una deviazione standard $\sigma = 5$.

Qual è il punteggio standardizzato corrispondente ai punteggi "grezzi" $x_1 = 65$ e $x_2 = 80$?

Si standardizzino i valori x

$$z_1 = \frac{65 - 70}{5} = -1$$

$$z_i = \frac{80 - 70}{5} = 2$$

$x_1 = 65$ è 1 deviazione standard sotto la media

$x_2 = 80$ è 2 deviazioni standard sopra la media

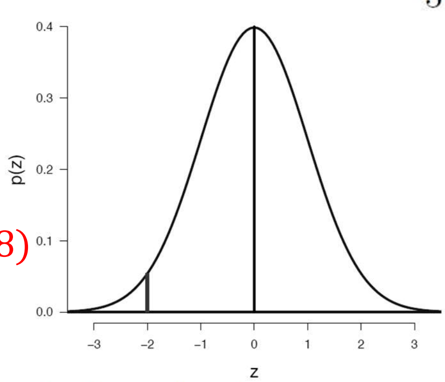
Illustrazione

Che percentuale di studenti riceve un punteggio minore di 60?

Si standardizzi il valore di $x = 60$

$$z_x = \frac{60 - 70}{5} = -2$$

Appendice del libro p. 286
$P(z \geq 2.00 = 0.0228)$



Dunque, circa il 2.3% degli studenti riceve un punteggio minore di 60.

Illustrazione

un milionario eccentrico offre un premio agli studenti che ottengono un punteggio maggiore di 83.75 nel test precedente.

Che percentuale di studenti otterrà il premio?

Si standardizzi il valore $x = 83.75$

$$z_1 = \frac{83.75 - 70}{5} = 2.75$$

Appendice del libro
p. 286
$P(z \geq 2.75 = 0.0029)$ ***Non 0.0020!***

Dunque, soltanto 3 studenti su 1000 vinceranno il premio.

Illustrazione

che percentuale di studenti ottiene punteggi compresi tra 70 e 80?

Si standardizzino i valori x

$$z_\mu = \frac{70 - 70}{5} = 0$$

$$z_2 = \frac{80 - 70}{5} = 2$$

Appendice del libro p. 286
$P(z \geq 2.00 = 0.0228)$
$P(z \geq 0.00 = 0.5000)$

Quindi, l'area compresa tra 0 e 2 è $0.5 - 0.0228 = 0.4772$. In altri termini, circa il 48% degli studenti ottiene punteggi compresi tra 70 e 80.

Illustrazione

essendosi convinto che la sua offerta non è particolarmente generosa, il milionario eccentrico dell'esempio precendente decide di offrire un premio al 5% degli studenti coni voti più alti.

Che voto deve ottenere uno studente per ricevere il premio?

Il problema richiede il valore z_i tale per cui l'intervallo $[z; \infty]$ sottenda il 5% dell'area. Dalla tabella troviamo che $z_{0.05}$, ovvero il 95$esimo$ percentile di $N(0,1)$, è $z_{0.05} = 1.645$. Tra 1.64 e 1.65

Quindi, circa il 95% dell'area è sottesa alla curva di densità normale standardizzata nell'intervallo $[-\infty; 1.645]$.

286 *Appendice*

Tavola 1. Aree soggiacenti alla distribuzione normale standardizzata

z	0	1	2	3	4	5	6	7	8	9
0.0	5000	4960	4920	4880	4840	4801	4761	4721	4651	4641
0.1	4602	4562	4522	4483	4443	4404	4364	4325	4286	4246
0.2	4207	4168	4129	4090	4052	4013	3974	3936	3897	3859
0.3	3821	3783	3745	3707	3669	3632	3594	3557	3520	3483
0.4	3446	3409	3372	3336	3300	264	3228	3192	3156	3121
0.5	3085	3050	3015	2981	2946	2912	2877	2843	2810	2776
0.6	2742	2709	2676	2643	2611	2578	2546	2514	2482	2451
0.7	2420	2388	2358	2327	2296	2266	2236	2206	2177	2148
0.8	2119	2090	2061	2033	2004	1977	1949	1922	1894	1867
0.9	1841	1814	1788	1762	1736	1711	1685	1660	1635	1611
1.0	1587	1562	1539	1515	1492	1469	1446	1423	1401	1379
1.1	1357	1335	1314	1292	1271	1251	1230	1210	1190	1170
1.2	1151	1131	1112	1093	1075	1056	1038	1020	1003	0985
1.3	0968	0951	0934	0918	0901	0885	0869	0853	0838	0812
1.4	0808	0793	0778	0764	0649	0735	0721	0708	0694	0681
1.5	0668	0655	0643	0630	~~0618~~	~~0606~~	0594	0582	0571	0559
1.6	0548	0537	0516	0516	0505	0495	0485	0475	0465	0455
1.7	0446	0436	0427	0418	0409	0401	0392	0384	0375	0367
1.8	0359	0351	0344	0336	0329	0322	0314	0307	0301	0294
1.9	0287	0281	0274	0268	0262	0256	0250	0244	0239	0233
2.0	0228	0222	0217	0212	0207	0202	0197	0192	0188	0183

Infine, per trasformare il punteggio z_i nel punteggio grezzo, avremo

$$x_{0.05} = \mu + z_{0.05}\sigma = 70 + 1.645 \times 5 = 78.23$$