Distribuzioni di probabilità

Concetto di probabilità

Esistono diverse definizioni:

- Classica: Dato un evento A $P(A)$ è il rapporto tra il numero dei casi favorevoli all'evento e il numero dei casi possibili, purché questi ultimi siano tutti equiprobabili.[1]. È un po' un cane che si morde la coda allora...
- Frequentista: Dato un evento A, indichiamo come successi (n_A) il numero di volte in cui osserviamo A su un totale di N prove. Costruiamo la frequenza relativa di A: $F(A) = \frac{n_A}{N}$. La probabiltà dell'evento è data dal limite per N che tende ad infinito di $F(A)$, ovvero

$$\lim_{N\to\infty} \frac{n_A}{N}$$

anche questa definizione è criticabile, ad esempio per mandare N all'infinito ho bisogno di poter ripetere le prove infinite volte

[1] Questa definizione spesso attribuita a Pierre Simon Laplace e quindi anche detta definizione classica di Laplace

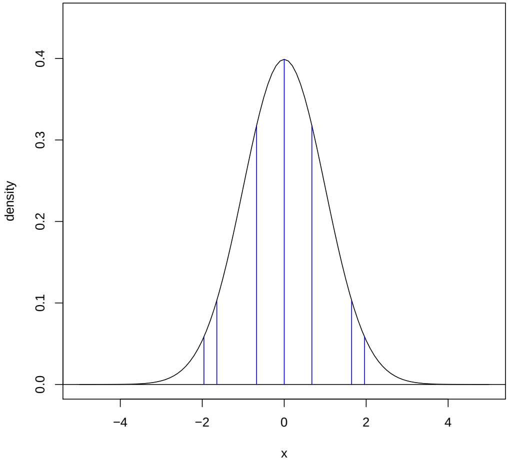
Praticamente...

Da tutte le definizioni però si ricavano tre regole/proprietà di base che devono essere vere affinché una funzione sia una probabilità:

1. $P(A)$ è compresa tra 0 e 1;
2. Se indichiamo con Ω l'insieme degli eventi possibili $P(\Omega) = 1$;
3. dati due eventi incompatibili A e B $(A \cap B = 0)$,
   $P(A \cup B) = P(A) + P(B)$.

- Un concetto molto importante in statistica è quello di *distribuzione*, spesso è un modo breve per dire *distribuzione di frequenza*
- Gli istogrammi mostrati in precedenza descrivono esattamente questo, sono istogrammi di distribuzioni empiriche cioè costruite su campioni
- Di solito cerchiamo di confrontare le distribuzioni empiriche con dei *modelli teorici* che ci permettano di costruire affermazioni rigorose sul comportamento dei dati.
- La distribuzione teorica più usata è la *normale* o *Gaussiana*

Normale di media 0 e varianza 1



Perché è tanto importante questa distribuzione?

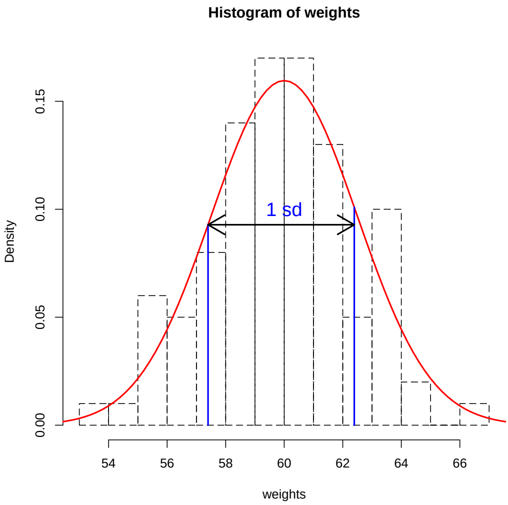
- Molti fenomeni naturali hanno questo comportamento simmetrico rispetto ad un valore e campanulare nella distribuzione di frequenza
- La si può spesso usare anche quando i dati non mostrano un andamento normale ma ricadono sotto le assunzioni del **teorema centrale del limite** che afferma che quando le osservazioni sono indipendenti ed hanno varianza finita, la media campionaria tende a distribuirsi come una normale al tendere ad infinito della dimensione del campione.

Perché è tanto importante questa distribuzione?

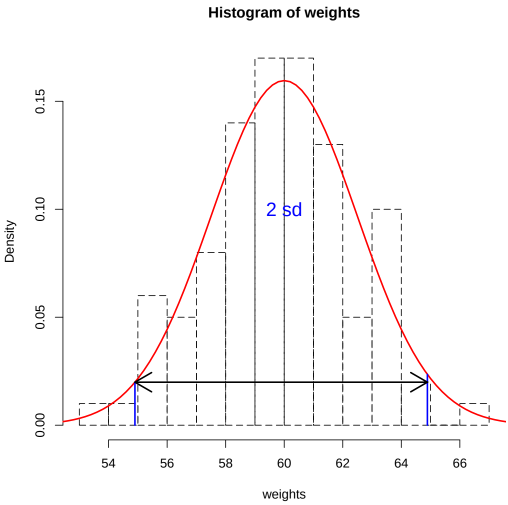
Regola empirica

- Prendiamo un insieme di osservazioni che segue la distribuzione normale, date media e varianza dei dati, stimate dal campione, sappiamo che circa
  1. il 68% dei dati cade in un intervallo di ampiezza 1 s.d. dalla media
  2. 95% dei dati cade in un intervallo di ampiezza 2 s.d. dalla media
  3. Quasi tutti i dati (99.7%) cadono in un intervallo di ampiezza 3 s.d. dalla media
- Attenzione: quanto detto è un'approssimazione, in termini rigorosi, ad esempio il 95% dei dati cade in un intervallo di 1.96 s.d.

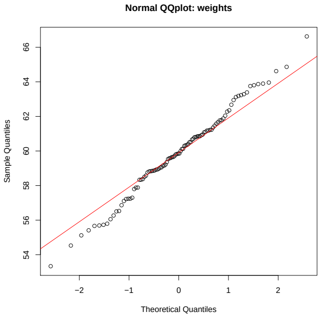
Esempio: pesi di 100 donne media=60kg, sd=2.5

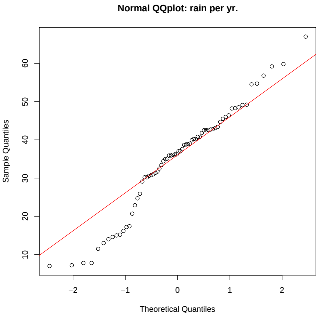


Esempio: pesi di 100 donne media=60kg, sd=2.5



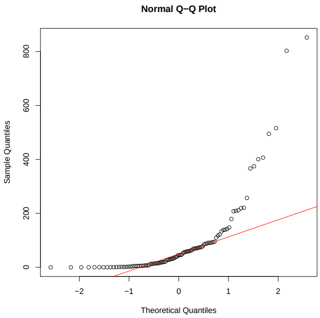
- Come capire se i dati seguono una distribuzione normale?
  1. QQ-plot: grafico a dispersione dei quantili empirici vs quelli teorici. Se i dati si dispongono lungo la bisettrice i dati seguono la distribuzione teorica
  2. istogramma
  3. test di adattamento (vedremo in seguito)

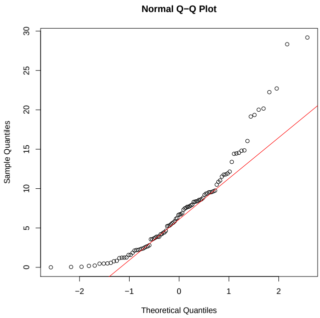




E se i dati non sono normali?

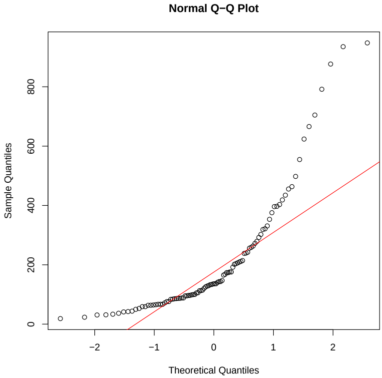
- Se i dati mostrano un comportamento molto lontano dalla normale possiamo sempre trasformarli
- le più usate sono la radice quadrata e il logaritmo naturale o in base 10 ...

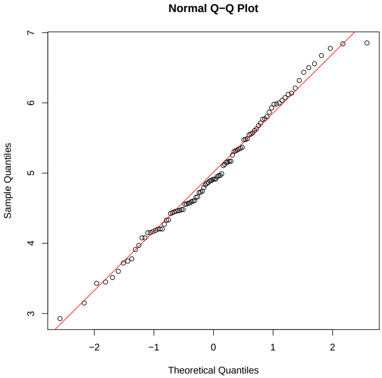




radice quadrata

E se i dati non sono normali?





logaritmo naturale

Alcune osservazioni

- Ora che abbiamo a disposizione la normale come modello teorico di riferimento possiamo porre delle domande più precise ai nostri dati
- Ad esempio possiamo chiedere, una volta stimata una media da un campione, quali siano i valori plausibili per quella media se generalizziamo il risultato
- Assumiamo che i dati siano normali, confermiamo questa ipotesi con un qq-plot e poi costruiamo un intervallo di valori attorno alla media che sia plausibile ad un certo livello di probabilità $(1-\alpha)$
- Formalmente

$$Prob(\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}) = 1-\alpha$$

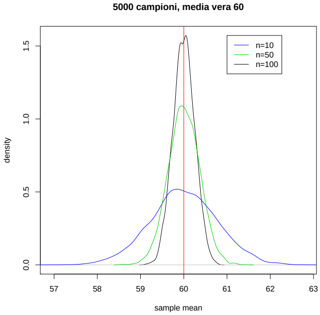
Intervallo di Confidenza

Alcune osservazioni

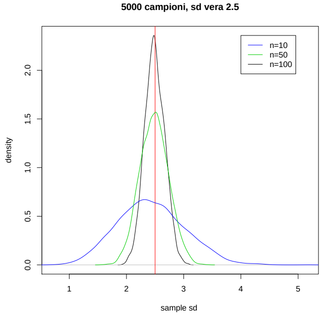
- Nella formula precedente $z_{1-\alpha/2} = 1, 1.96, 3$ etc, α è *il livello di confidenza* (0.318,0.05,0.002 rispettivamente) ed s è la stima della deviazione standard ottenuta dal campione.
- La normale la usiamo in tantissime occasione come modello teorico di riferimento
- Uno degli usi fondamentali è proprio la costruzione di *Intervalli di Confidenza*

Alcune osservazioni

- In generale, per identificare quale normale descriva bene i nostri dati usiamo la media e la varianza campionarie
- Bisogna ricordare che queste due quantità non sono oggetti fissati ma sono a loro volta variabili

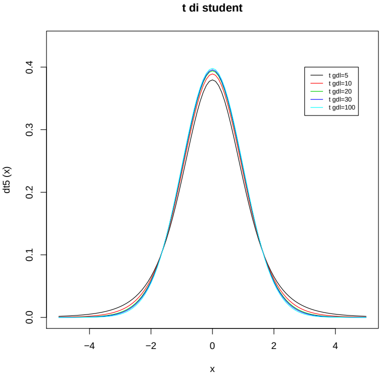


media campionaria al variare di n

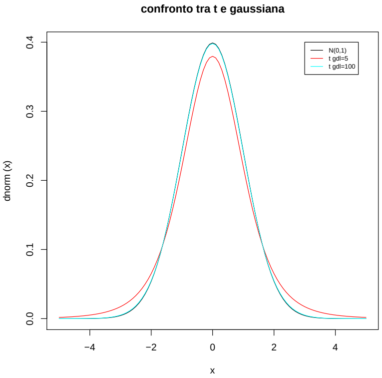


sd campionaria al variare di n

- Dai grafici vediamo chiaramente che se il campione è piccolo la varianza campionaria non fornisce una stima affidabile della “vera varianza.
- Per compensare questa incertezza (mantenendo lo stesso livello di confidenza) che aumenta al diminuire delle dimensioni del campione, si usa la distribuzione t che dipende da n.



Confronto tra Normale e t



- Quindi i nostri intervalli di confidenza ora saranno

$$\bar{x} \pm t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}$$

- con t valore (quantile) da determinare di volta in volta sulla base di n e del livello di confidenza scelto

Quando usare la t invece della normale

Regola Empirica

- **Numerosità del campione inferiore a 15**: Si usa la distribuzione t solo se la distribuzione dei dati è simile a una normale (nessun outlier - valore anomalo)
- **Numerosità del campione superiore a 15** Si fa riferimento alla distribuzione t, tranne in presenza di outlier o forte asimmetria
- **Grandi campioni ($n \geq 40$)** Si può usare la t anche nel caso di distribuzioni molto asimmetriche

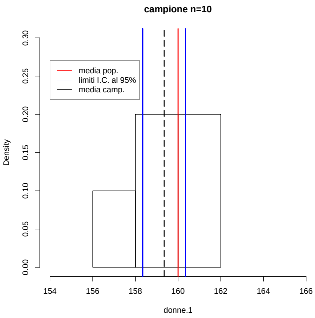
Esempio

- Prendiamo due campioni di misure di altezza di donne da una popolazione che ha la stessa media e la stessa varianza. Un campione con $n = 10$ e l'altro con $n = 50$ osservazioni

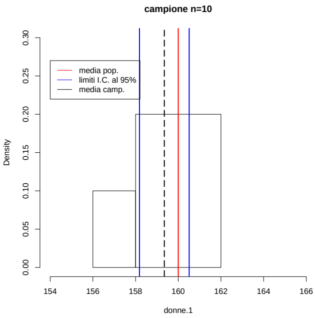
valori	$n = 10$	$n = 50$
media	159.35	160.13
sd della media	0.51	0.28
IC normale	[158.34, 160.35]	[159.59, 160.67]
IC t	[158.18, 160.51]	[158.31, 160.38]

Esempio

Campione di 10 donne



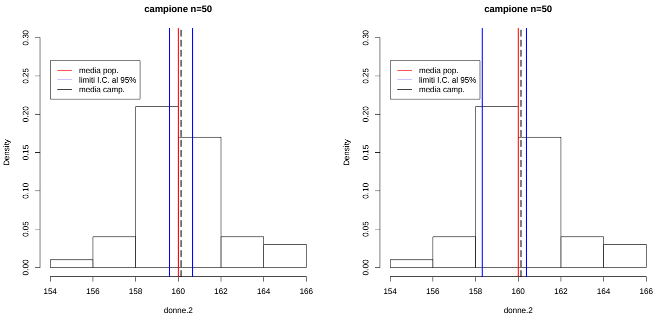
I.C. normale



I.C. t-student 9 gdl

Esempio

Campione di 50 donne



I.C. normale I.C. t-student 9 gdl