Statistica e Laboratorio

6. Calcolo delle probabilità: modelli probabilistici

Paolo Vidoni

Dipartimento di Scienze Economiche e Statistiche
Università di Udine
via Tomadini 30/a - Udine
paolo.vidoni@uniud.it

https://elearning.uniud.it/

Sommario

Sommario

- Introduzione
- Modelli discreti
- Modelli continui

Introduzione

Per **modello probabilistico** si intende non una singola distribuzione di probabilità, ma una famiglia di distribuzioni di probabilità.

In genere, le distribuzioni di probabilità appartenenti ad una determinata famiglia presentano funzione di densità (probabilità) con la medesima forma funzionale e vengono individuate in corrispondenza ai diversi valori assunti da una o più costanti reali chiamate **parametri**.

Data un variabile casuale X, verrà definito il supporto S_X e la famiglia di funzioni di densità (probabilità) $f_X(x;\theta)$, $\theta \in \Theta \subseteq \mathbf{R}^n$, $n \geq 1$, dove θ è il parametro (vettore dei parametri) che individua le distribuzioni di probabilità appartenenti alla famiglia in esame.

Le **famiglie di distribuzioni di probabilità** (modelli) che verranno presentate sono utili per descrivere esperimenti aleatori tipici e sono individuate da un nome che le caratterizza.

In molti casi, tale terminologia viene trasferita alla variabile casuale che presenta come distribuzione di probabilità una di quelle della famiglia corrispondente.

Sommario

Modello uniforme discreto

Il modello uniforme discreto descrive esperimenti con un *numero finito di esiti equiprobabili*.

Una variabile casuale X ha distribuzione **uniforme discreta** con possibili valori $x_1, \ldots, x_n \in \mathbf{R}$, $n \in \mathbf{N}^+$ fissato, in simboli $X \sim Ud(x_1, \ldots, x_n)$, se $S_X = \{x_1, \ldots, x_n\}$ e

$$f_X(x; x_1, \ldots, x_n) = \begin{cases} 1/n & \text{se } x = x_1, \ldots, x_n \\ 0 & \text{altrimenti} \end{cases}$$

Inoltre, $E(X) = \sum_{i=1}^{n} x_i/n$, $V(X) = \sum_{i=1}^{n} (x_i - E(X))^2/n$.

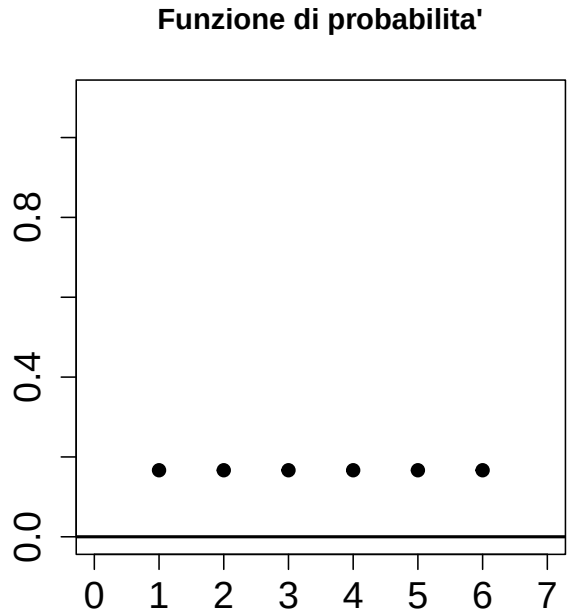
Se $x_i = i$, $i = 1, \ldots, n$, si scrive $X \sim Ud(n)$ e

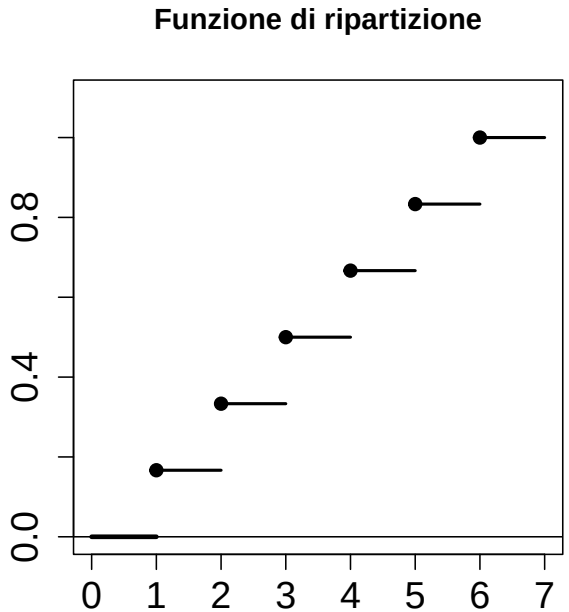
$$E(X) = \frac{n+1}{2}, \quad V(X) = \frac{n^2 - 1}{12}.$$

Se, in particolare, $n = 1$, si ottiene la distribuzione **degenere** in x_1, in simboli $X \sim D(x_1)$. In questo caso, $E(X) = x_1$ e $V(X) = 0$.

Esempio. *Dado*. Si consideri il lancio di un dado regolare. La variabile casuale X, che indica la faccia uscita dopo il lancio, ha distribuzione di probabilità $Ud(6)$.

Le figure seguenti rappresentano le associate funzioni di probabilità e di ripartizione di X.





$\Diamond$

Modello binomiale

Si considerano esperimenti che possono essere rappresentati come estrazioni con reinserimento da un'urna di composizione nota.

Ogni estrazione può essere classificata in due categorie incompatibili ed esaustive chiamate, in modo convenzionale, **successo** e **insuccesso** (osservazioni dicotomiche dove, in genere, 1 indica il successo e 0 l'insuccesso): **esperimento bernoulliano**.

Ogni estrazione è indipendente dalle altre e presenta la stessa probabilità $p \in (0, 1)$ di successo.

Il modello binomiale descrive il *numero di successi in* $n \geq 1$ *esperimenti bernoulliani indipendenti con la stessa probabilità di successo* $p \in (0, 1)$.

Una applicazione possibile è al *controllo di qualità*: si è interessati al numero di elementi difettosi in un campione casuale di dimensione $n \geq 1$, con $p \in (0, 1)$ la porzione di elementi difettosi.

Un'altra applicazione è al contesto delle *indagini di mercato*: si è interessati al numero di consumatori che apprezzano un certo prodotto in un campione casuale di dimensione $n \geq 1$, con $p \in (0, 1)$ la porzione di individui che apprezzano il prodotto.

Una ulteriore applicazione è allo *studio delle popolazioni*: si è interessati al numero di individui che presentano un certa caratteristica in un campione casuale di dimensione $n \geq 1$, con $p \in (0, 1)$ la porzione di individui portatori della caratteristica.

Se, come spesso accade nel campionamento da popolazione finita, si effettuano estrazioni senza reinserimento (estrazione in blocco), si può comunque utilizzare il modello binomiale se la popolazione è così elevata da essere considerata *quasi infinita*.

In questo caso, ha poca importanza se l'estrazione è fatta con o senza reinserimento.

Una variabile casuale X ha distribuzione **binomiale** di parametri $n \geq 1$ e $p \in (0,1)$, in simboli $X \sim Bi(n,p)$, se $S_X = \{0, \ldots, n\}$ e

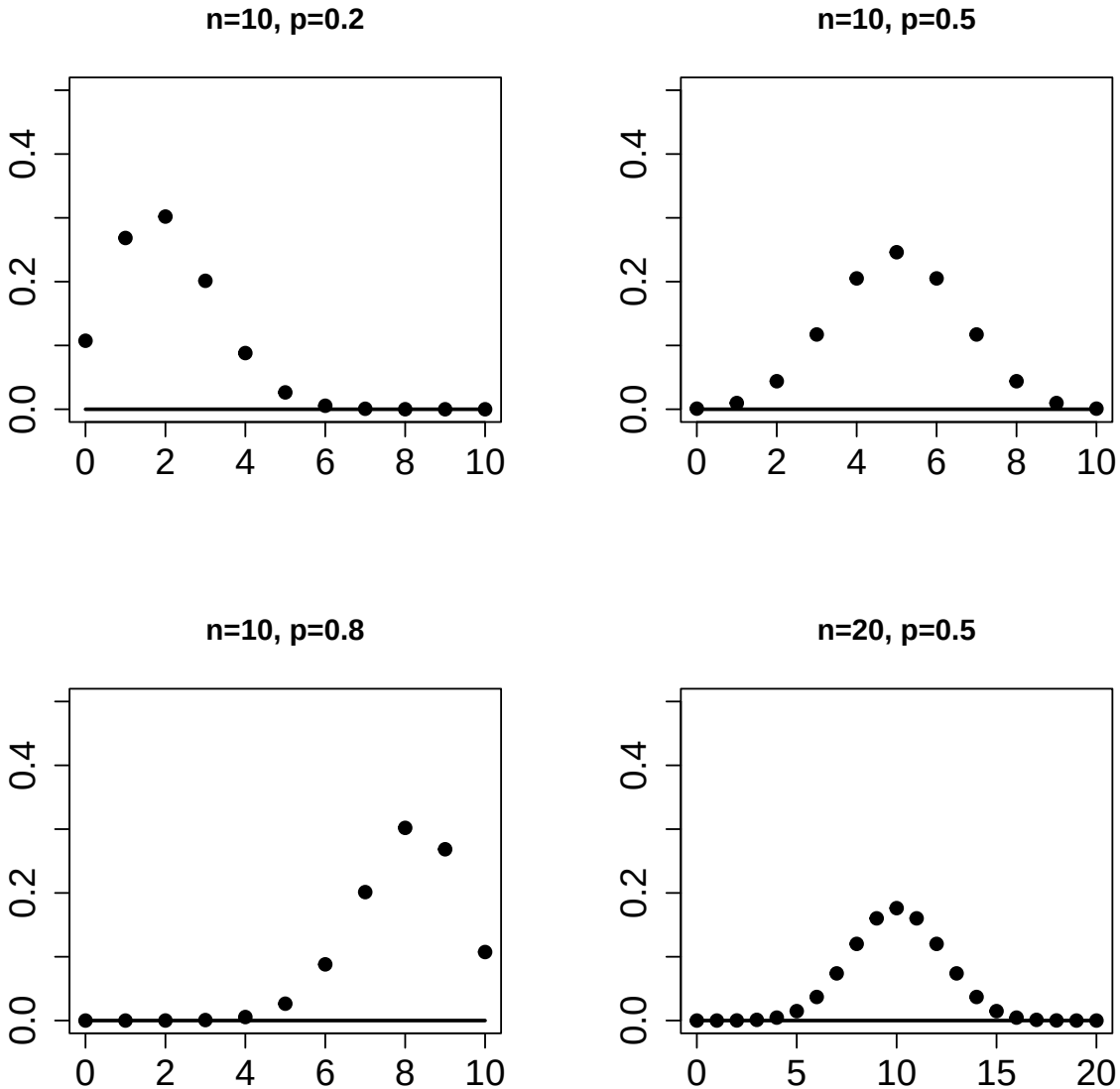
$$f_X(x; n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{se } x \in S_X \\ 0 & \text{altrimenti} \end{cases}$$

dove n indica il numero di prove (esperimenti bernoulliani) indipendenti e p la comune probabilità di successo.

È chiaro che $p^x(1-p)^{n-x}$ indica la probabilità di osservare x successi e $n-x$ insuccessi, in una specifica configurazione, e il coefficiente binomiale individua il numero di possibili configurazioni con x successi.

Se $n = 1$ si ha una variabile casuale **bernoulliana**, o **binomiale elementare**, in simboli $Ber(p)$ o $Bi(1,p)$.

Si considerano i grafici delle funzioni di probabilità nel caso in cui $n = 10$ e $p = 0.2, 0.5, 0.8$ e $n = 20$ e $p = 0.5$.



Se le variabili casuali $X_i \sim Ber(p)$, $i = 1, \ldots, n$, descrivono n esperimenti bernoulliani indipendenti, si può concludere che la variabile casuale somma $X = \sum_{i=1}^{n} X_i \sim Bi(n, p)$.

Si verifica facilmente che, per ogni $i = 1, \ldots, n$,

$$E(X_i) = 1 \cdot p + 0 \cdot (1 - p) = p,$$

$$V(X_i) = E(X_i^2) - (E(X_i))^2 = p(1 - p).$$

Quindi,

$$E(X) = E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) = np,$$

$$V(X) = V(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} V(X_i) = np(1 - p).$$

Infine, è facile verificare che la frequenza campionaria di successo (media campionaria di variabili bernoulliane) $Y = X/n = \sum_{i=1}^{n} X_i/n$ è tale che $E(Y) = p$ e $V(Y) = p(1 - p)/n$.

Esempio. *Atleti*. Tra i 100 iscritti ad una associazione sportiva ci sono 30 più alti di 180 cm. Si estrae casualmente un campione di $n = 10$ atleti con reinserimento.

La variabile casuale X che definisce il numero di atleti che, tra i 10 considerati, è più alto di 180 cm (*successo*) ha distribuzione $Bi(10, 0.3)$. Ci si attende di osservare $E(X) = 3$ atleti con altezza superiore a 180 cm ed inoltre $V(X) = 2.1$

La probabilità di estrarre almeno un atleta più alto di 180 cm è

$$P(X \geq 1) = 1 - P(X = 0) = 1 - [10!/(0!10!)]0.3^0(1 - 0.3)^{10} = 0.97.$$

La probabilità di estrarre due atleti più alti di 180 cm è

$$P(X = 2) = [10!/(2!8!)]0.3^2(1 - 0.3)^8 = 0.23.$$

Infine, la probabilità di estrarne meno di 4 è

$$P(X < 4) = \sum_{i=0}^{3} P(X = x_i) = 0.27 + 0.23 + 0.12 + 0.03 = 0.65.$$

$\Diamond$

Esempio. *Monitor*. Per un inconveniente nella linea di produzione, su 100 monitor prodotti da una certa azienda 10 risultano difettosi. Un rivenditore ha, recentemente, acquistato cinquanta monitor da questa azienda.

La variabile casuale X che descrive il numero di monitor che, tra i cinquanta venduti, verranno resi alla casa produttrice perché difettosi è una $Bi(50, 0.1)$.

Il numero atteso di monitor difettosi è $E(X) = 5$, mentre $V(X) = 4.5$. Inoltre, la probabilità che nessun monitor sia difettoso è

$$P(X = 0) = [50!/(0!50!)]0.1^0(1 - 0.1)^{50} = 0.005$$

$\Diamond$

Modello poisson

Il modello poisson descrive *problemi di conteggio* quando non c'è una limitazione superiore per il supporto o problemi in cui tale limitazione è praticamente irrilevante.

Sotto alcune ipotesi, descrive il *numero di arrivi o accadimenti* di un evento di interesse (*successo*) in un intervallo di tempo (o anche su una superficie) di dimensione fissata.

Una variabile casuale X ha distribuzione **poisson** con parametro $\lambda > 0$, in simboli $X \sim P(\lambda)$, se $S_X = \mathbf{N}$ e

$$f_X(x;\lambda) = \begin{cases} \lambda^x e^{-\lambda}/x! & \text{se } x \in S_X \\ 0 & \text{altrimenti} \end{cases}$$

Si dimostra che $E(X) = \lambda$ e $E(X^2) = \lambda^2 + \lambda$, da cui si ottiene che $V(X) = E(X^2) - (E(X))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$. Quindi, media e varianza coincidono e corrispondono al parametro λ.

Se il numero medio di *successi* in una unità di tempo è ν, la variabile casuale che rappresenta il numero di successi in un intervallo di tempo t ha distribuzione $P(\lambda)$, con $\lambda = \nu t$.

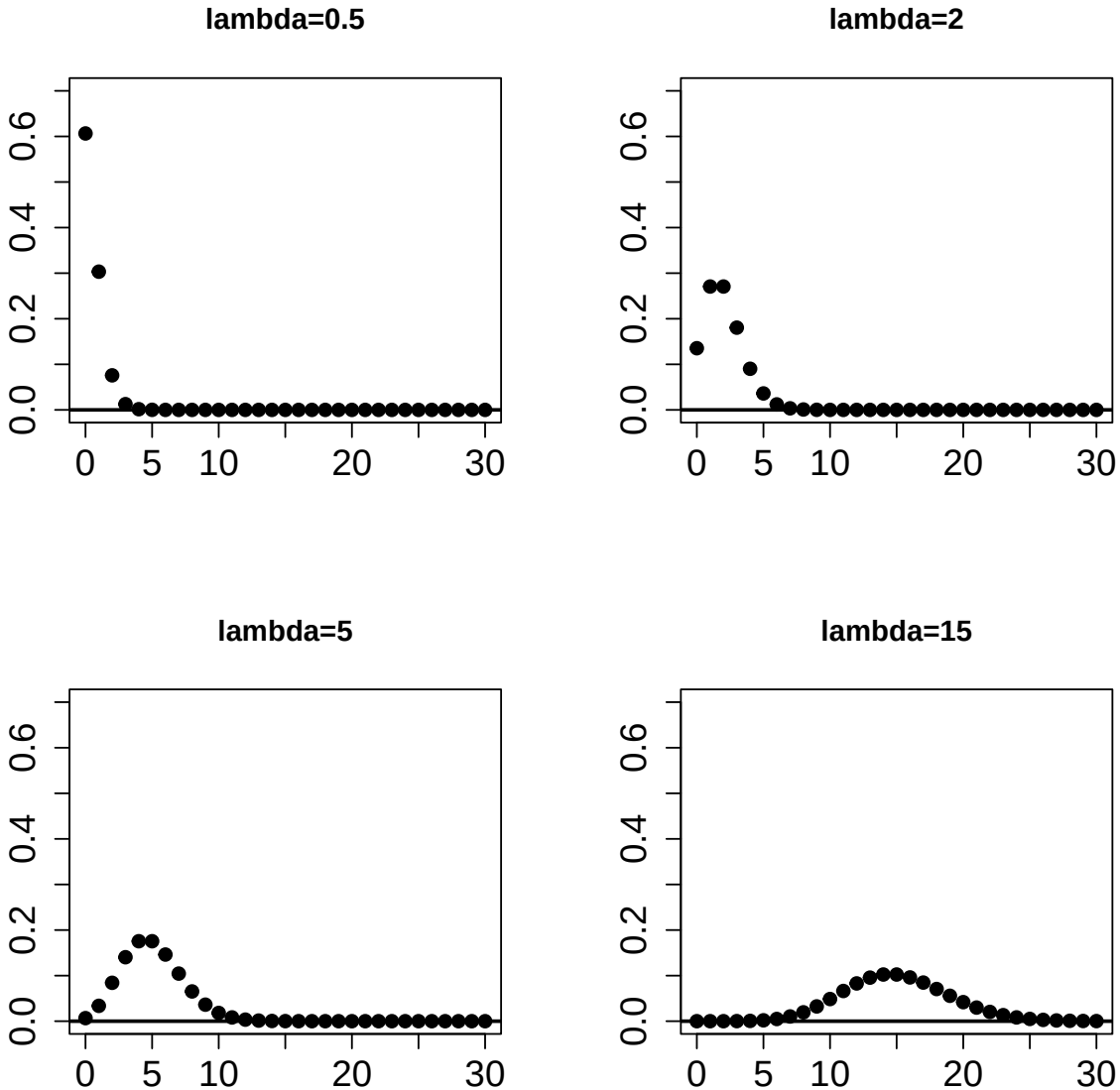
La distribuzione poisson può venire interpretata come caso limite della distribuzione binomiale. Se n è elevato e p è piccola, la distribuzione $Bi(n, p)$ viene approssimata da una $P(\lambda)$, con $\lambda = np$.

Per quanto riguarda le applicazioni, l'approssimazione è efficace se $n \geq 50$ e $p \leq 1/25$.

L'approssimazione risulta pressoché esatta quando si considerano fenomeni come, ad esempio, il numero di cittadini americani coinvolti in incidenti stradali in un anno, dove $n = 303824646$ (luglio 2008, stima) e $p = 0.00024$.

Se $Y_1 \sim P(\lambda_1)$ e $Y_2 \sim P(\lambda_2)$ sono indipendenti (gli eventi associati a una variabile sono indipendenti da quelli associati all'altra), la loro somma $Y_1 + Y_2 \sim P(\lambda_1 + \lambda_2)$.

Si considerano i grafici delle funzioni di probabilità nel caso in cui $\lambda = 0.5, 2, 5, 15$.



Esempio. *Pronto soccorso*. Al Pronto Soccorso di un piccolo ospedale si presentano in media 3 pazienti ogni ora.

Per predisporre il personale medico necessario, si vuole calcolare la probabilità che in un'ora arrivino esattamente 2 pazienti e la probabilità che in un'ora arrivino più di 2 pazienti.

Indicata con $X \sim P(3)$ la variabile casuale che descrive il numero di arrivi in un'ora al Pronto Soccorso, la probabilità che arrivino 2 pazienti in un'ora è

$$P(X = 2) = \frac{e^{-3}3^2}{2!} = 0.224,$$

mentre la probabilità che arrivino più di 2 pazienti in un'ora è

$$P(X > 2) = 1 - P(X \leq 2) = 1 - \sum_{i=0}^{2} P(X = i) = 1 - 0.423 = 0.577.$$

$\Diamond$

Esempio. *Pezzi difettosi*. Un certo macchinario produce un pezzo difettoso ogni cento. Si vuole calcolare la probabilità che, scegliendo a caso 100 pezzi, se ne trovino esattamente 3 difettosi.

La probabilità esatta, ottenuta utilizzando la distribuzione $Bi(100, 0.01)$, è

$$\begin{pmatrix} 100 \\ 3 \end{pmatrix} 0.01^3 0.99^{97} = 0.0609.$$

Essendo n elevato e p piccolo, si può utilizzare l'approssimazione con la distribuzione $P(\lambda)$, con $\lambda = np = 1$. Si ottiene

$$\frac{e^{-1}1^3}{3!} \doteq 0.0613,$$

valore prossimo a quello esatto. $\Diamond$

Modello geometrico

Il modello geometrico descrive il *tempo di attesa*, espresso come numero di replicazioni indipendenti di un esperimento bernoulliano, con probabilità di successo p, per osservare per la prima volta un successo.

Una variabile casuale X ha distribuzione **geometrica** con parametro $p \in (0, 1)$, in simboli $X \sim Ge(p)$, se $S_X = \mathbf{N}^+$ e

$$f_X(x; \lambda) = \begin{cases} (1-p)^{x-1}p & \text{se } x \in S_X \\ 0 & \text{altrimenti} \end{cases}$$

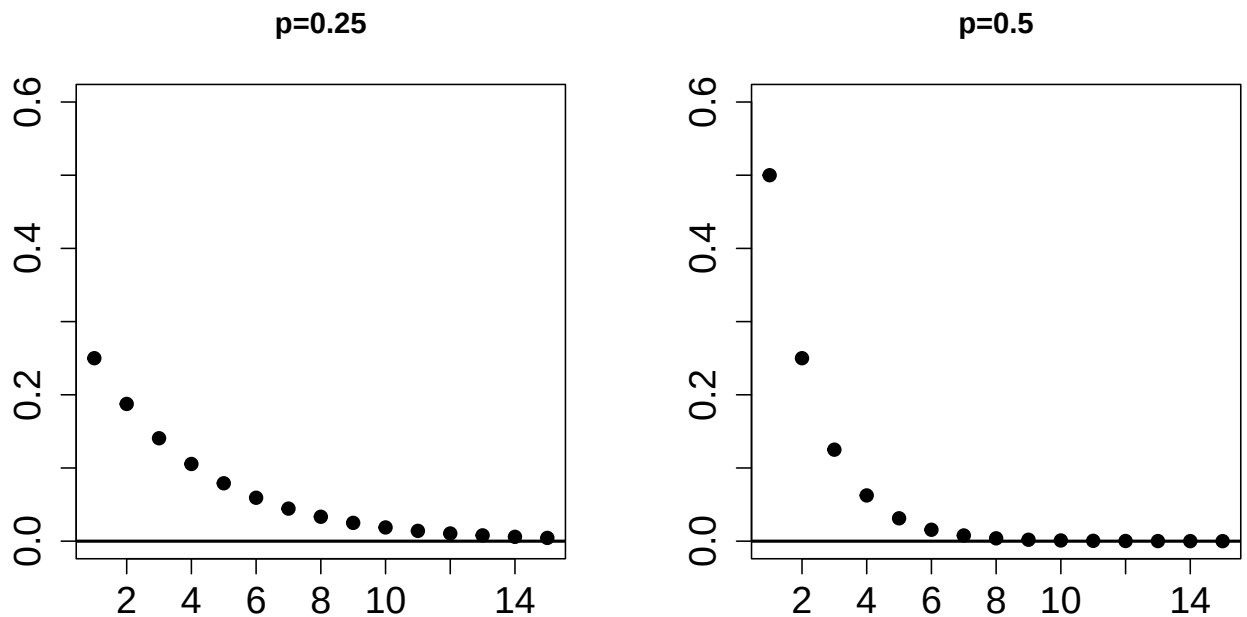
Si dimostra che $E(X) = 1/p$ e $V(X) = (1-p)/p^2$.

Il modello geometrico è caratterizzato dalla proprietà di **assenza di memoria**, che viene specificata dalla seguente condizione

$$P(X > s + t | X > s) = P(X > t), \text{ per ogni } s, t \in S_X.$$

Quindi la probabilità che il successo avvenga dopo $s + t$ prove bernoulliane, sotto la condizione che non sia avvenuto nelle prime s prove, pari alla probabilità non condizionata che il successo avvenga dopo t prove.

Si considerano i grafici delle funzioni di probabilità nel caso in cui $p = 0.25, 0.5$.



Esempio. *Gioco del lotto*. Si consideri il gioco del lotto. La probabilità che esca il tre in una singola estrazione su una ruota prefissata $1/18$.

La variabile casuale X, che indica il numero di settimane necessarie affinché esca il numero tre sulla ruota di Napoli, ha distribuzione $Ge(1/18)$.

Si calcola la probabilità che il tre esca alla trentesima settimana, se si è a conoscenza che non è uscito nelle prime dieci settimane

$$P(X = 30|X > 10) = \frac{P(X = 30)}{P(X > 10)} = \frac{(17/18)^{29}(1/18)}{(17/18)^{10}} = \left(\frac{17}{18}\right)^{19}\frac{1}{18}.$$

La probabilità cercata corrisponde a $P(X = 20)$, cioè alla probabilità non condizionata che il tre esca alla ventesima settimana.

Questo risultato conferma la totale infondatezza dell'uso dei ritardi per congetturare una modificazione delle probabilità di successo nel gioco del lotto. $\Diamond$

Sommario

Modello uniforme continuo

Il modello uniforme continuo descrive esperimenti aleatori che possono essere rappresentati come un'estrazione casuale di un numero dall'intervallo $[a, b]$.

Il concetto di *equiprobabilità* viene trasferito all'*ambito continuo* richiedendo che tutti i sotto-intervalli del supporto di uguale lunghezza abbiano la stessa probabilità di contenere il risultato sperimentale.

Una variabile casuale X ha distribuzione **uniforme continua** (**rettangolare**) con parametri $a, b \in \mathbf{R}$, $a < b$, in simboli $X \sim U(a, b)$, se $S_X = [a, b]$ e

$$f_X(x; a, b) = \begin{cases} 1/(b-a) & \text{se } a \leq x \leq b \\ 0 & \text{altrimenti} \end{cases}$$

$$F_X(x; a, b) = \begin{cases} 0 & \text{se } x < a \\ (x-a)/(b-a) & \text{se } a \leq x < b \\ 1 & \text{se } x \geq b. \end{cases}$$

Si verifica facilmente che

$$E(X) = \int_a^b x \frac{1}{b-a} dx = \frac{b^2-a^2}{2} \frac{1}{b-a} = \frac{b+a}{2},$$

$$\begin{aligned} V(X) &= E(X^2) - (E(X))^2 = \int_a^b x^2 \frac{1}{b-a} dx - \left(\frac{b+a}{2}\right)^2 \\ &= \frac{b^3-a^3}{3(b-a)} - \left(\frac{b+a}{2}\right)^2 = \frac{(b-a)^2}{12}. \end{aligned}$$

Per il grafico delle funzioni di densità e di ripartizione si rimanda all'esempio presentato in precedenza, dove si è considerata una variabile casuale $U(0,1)$, dove $a = 0$ e $b = 1$.

La distribuzione uniforme continua chiusa rispetto alle trasformazioni di posizione e di scala. Infatti, se $X \sim U(a, b)$, allora $Y = \alpha + \beta X \sim U(\alpha + \beta a, \alpha + \beta b)$, con $\alpha, \beta \in \mathbf{R}$.

Esempio. *Criceti*. Si predispone un esperimento per valutare il senso di orientamento dei criceti. Gli animali vengono posti al centro un contenitore circolare con un'unica via di uscita. Dopo averli bendati e disorientati, si osserva la direzione scelta da ciascun criceto.

Sia $X \sim U(-\pi, \pi)$ la variabile casuale che esprime l'ampiezza in radianti dell'angolo tra la direzione scelta dall'animale e la direzione che porta all'uscita. Si ha che $E(X) = 0$ e $V(X) = \pi^2/3$.

Si consideri la variabile casuale Y, che esprime l'ampiezza dell'angolo in gradi. Poiché $Y = 180X/\pi$, si ha che $Y \sim U(-180, 180)$. $\Diamond$

Modello esponenziale

Il modello esponenziale viene utilizzato soprattutto per rappresentare *durate* e *tempi di vita o di funzionamento* (ad esempio negli studi di affidabilità), nel caso in cui sia plausibile assumere la proprietà di *assenza di memoria o di usura*.

Una variabile casuale X ha distribuzione **esponenziale** con parametro $\lambda > 0$, detto **tasso di guasto**, in simboli $X \sim Esp(\lambda)$, se $S_X = [0, +\infty)$ e

$$f_X(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{se } x \in S_X \\ 0 & \text{altrimenti} \end{cases}$$

$$F_X(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & \text{se } x > 0 \\ 0 & \text{se } x \leq 0. \end{cases}$$

In alcuni casi si utilizza la parametrizzazione alternativa con $\theta = 1/\lambda$, che corrisponde al valor medio di X.

Nell'esempio presentato in precedenza si è considerato il grafico delle funzioni di densità e di ripartizione per il caso $\lambda = 1$ e si è dimostrato che

$$E(X) = \frac{1}{\lambda}, \quad V(X) = \frac{1}{\lambda^2}.$$

Inoltre, vale la seguente proprietà di chiusura rispetto a trasformazioni di scala: se $X \sim Esp(\lambda)$, si ha che $aX \sim Esp(\lambda/a)$, per ogni $a > 0$.

La proprietà di **assenza di memoria** (**usura**) caratterizza la distribuzione esponenziale ed è specificata dalla relazione

$$P(X > s + t | X > t) = P(X > s),$$

per ogni $s, t > 0$, che l'analogo nel continuo della condizione specificata, nel caso discreto, per la distribuzione geometrica.

Esempio. *Circuito*. Un circuito è costituito da due componenti dal funzionamento indipendente la cui vita operativa, misurata in anni, è descritta rispettivamente dalle variabili casuali $X_1 \sim Esp(0.2)$ e $X_2 \sim Esp(0.3)$.

Si cerca la probabilità che funzionamento del circuito sia non superiore a 10 anni. Se i componenti sono in parallelo, si ha

$$\begin{aligned} P(X_1 \leq 10 \cap X_2 \leq 10) &= P(X_1 \leq 10)P(X_2 \leq 10) \\ &= (1 - e^{-0.2 \cdot 10})(1 - e^{-0.3 \cdot 10}) = 0.822. \end{aligned}$$

Se i componenti sono in serie, si ha

$$\begin{aligned} 1 - P(X_1 > 10 \cap X_2 > 10) &= 1 - P(X_1 > 10)P(X_2 > 10) \\ &= 1 - e^{-0.2 \cdot 10} e^{-0.3 \cdot 10} = 0.993. \end{aligned}$$

In entrambi i casi si è tenuto conto che gli eventi riferiti a X_1 sono indipendenti da quelli riferiti a X_2. $\Diamond$

Modello normale

Il modello normale o gaussiano è il *modello più importante* ed è anche il *più utilizzato* nelle applicazioni della statistica inferenziale.

È stato studiato, in particolare, dal matematico tedesco K.F. Gauss, da cui il termine gaussiano, che nel 1809 lo utilizzò per descrivere gli *errori accidentali* (non sistematici), nel caso di misurazioni strumentali ripetute di una grandezza incognita.

Viene utilizzato in vari contesti di applicazione. In particolare risulta utile per descrivere, oltre agli errori accidentali, la presenza di *caratteri antropometrici*, come la statura e il peso, in popolazioni umane omogenee, oppure per studiare alcuni *particolari fenomeni sociali o naturali*.

Infine, viene utilizzato in molte applicazione come *modello per approssimare*, con buona accuratezza, diverse distribuzioni di probabilità, discrete e continue, e permette quindi di agevolare i calcoli.

Una variabile casuale X ha distribuzione **normale** o **gaussiana** con parametri $\mu \in \mathbf{R}$ e $\sigma^2 > 0$, in simboli $X \sim N(\mu, \sigma^2)$, se $S_X = \mathbf{R}$ e, per ogni $x \in \mathbf{R}$,

$$f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

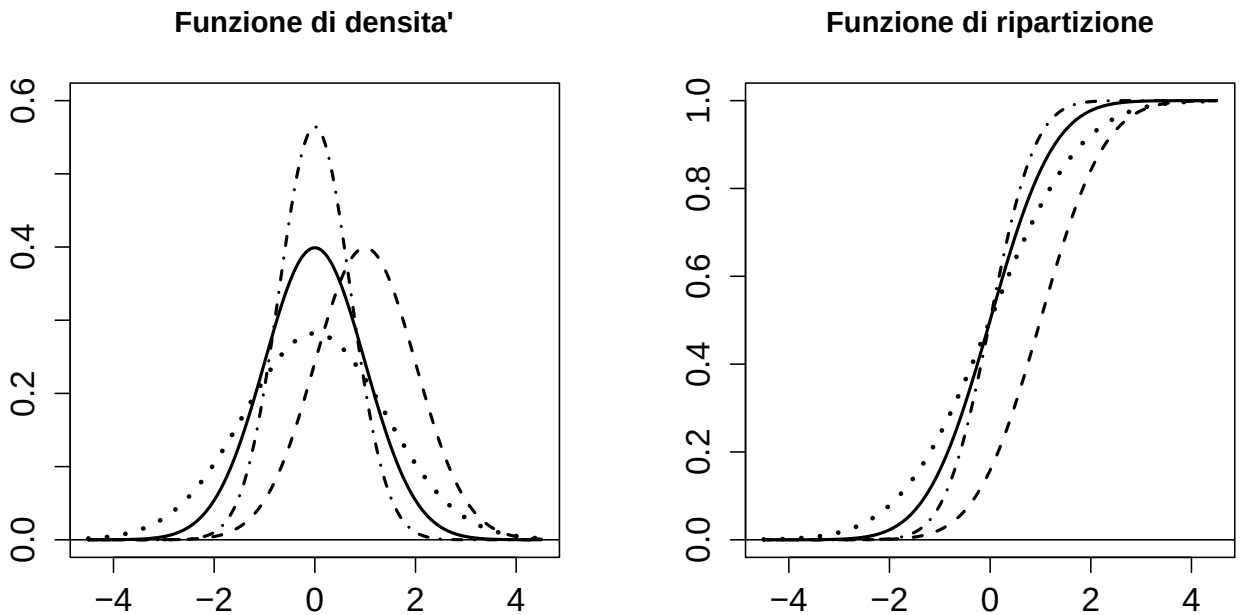
La corrispondente funzione di ripartizione non ha una forma esplicita e viene definita implicitamente utilizzando la definizione di funzione di ripartizione.

Come risulta chiaro dai grafici presentati nel seguito, la funzione $f_X(x; \mu, \sigma)$ ha un massimo assoluto nel punto μ e punti di flesso nei punti $\mu - \sigma$ e $\mu + \sigma$. Inoltre, risulta simmetrica rispetto alla retta $x = \mu$.

Quindi, il parametro μ è sia la moda che la mediana. Inoltre, si verifica che $E(X) = \mu$ e $V(X) = \sigma^2$.

Se $\mu = 0$ e $\sigma^2 = 1$ si ottiene la distribuzione **normale standard**, in simboli $N(0, 1)$.

Si riporta il grafico delle funzioni di densità e di ripartizione della variabile casuale $X \sim N(\mu, \sigma^2)$ per $\mu = 0$, $\sigma^2 = 1$ (—), $\mu = 1$, $\sigma^2 = 1$ (– –), $\mu = 0$, $\sigma^2 = 2$ $(\cdots)$, $\mu = 0$, $\sigma^2 = 1/2$ (- · -).



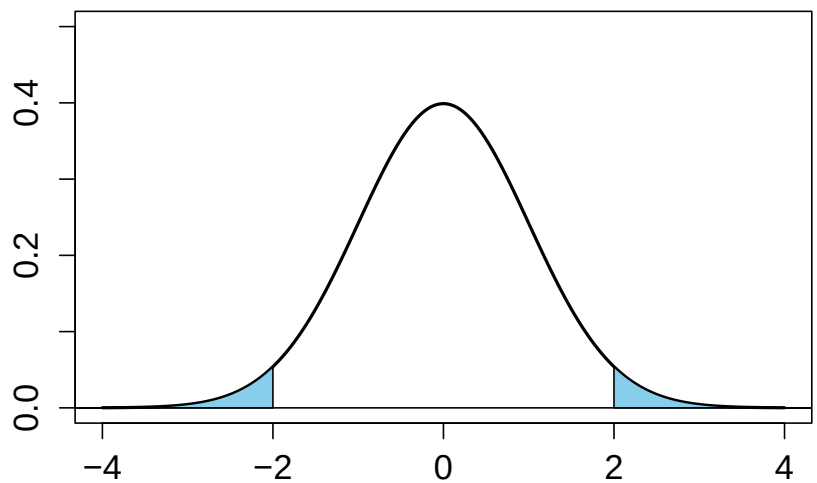
La distribuzione normale è chiusa rispetto alle trasformazioni lineari.

Infatti, si può verificare che, se $X \sim N(\mu, \sigma^2)$ e $Y = aX + b$, con $a, b \in \mathbf{R}$, $a \neq 0$, allora $Y \sim N(a\mu + b, a^2\sigma^2)$.

In particolare, se $X \sim N(\mu, \sigma^2)$, si ottiene una $Z \sim N(0, 1)$ con l'operazione di **standardizzazione** $Z = (X - \mu)/\sigma$. Viceversa, se $Z \sim N(0, 1)$, allora $X = \sigma X + \mu \sim N(\mu, \sigma^2)$.

La funzione di ripartizione e la funzione di densità di una $Z \sim N(0, 1)$ si indicano con $\Phi(z)$ e $\phi(z)$, rispettivamente. Poiché $\phi(x)$ è simmetrica rispetto all'origine, si ha che

$$\Phi(-z) = 1 - \Phi(z), \quad \forall z \geq 0.$$



Inoltre, dalla analisi del grafico si conclude che

$$P(|\ Z\ | < z) = \Phi(z) - \Phi(-z), \quad P(|\ Z\ | > z) = 2(1 - \Phi(z)).$$

Con la standardizzazione si elimina la dipendenza da particolari valori di μ e σ^2 e ci si riconduce ad una normale standard. Quindi, il calcolo di probabilità riferite ad una $X \sim N(\mu, \sigma^2)$ si traduce nel calcolo di probabilità di opportuni eventi associati a $Z \sim N(0, 1)$.

In particolare, per ogni $a, b \in \mathbf{R}$, $a < b$,

$$\begin{aligned} P(a \leq X \leq b) &= P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right), \end{aligned}$$

$$P(X \leq b) = F_X(b; \mu, \sigma) = P\left(\frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right).$$

Con una semplice applicazione dei risultati esposti, si ha che

$$\begin{aligned} P(\mu - \sigma < X < \mu + \sigma) &\doteq 0.68, \\ P(\mu - 2\sigma < X < \mu + 2\sigma) &\doteq 0.95, \\ P(\mu - 3\sigma < X < \mu + 3\sigma) &\doteq 0.997. \end{aligned}$$

La maggior parte della massa di probabilità si trova entro i limiti definiti dalla media più o meno 3σ (*regola dei 6 sigma*).

Utilizzando le **tavole statistiche** si ottengono facilmente i valori della funzione di ripartizione $\Phi(z)$ di una $N(0,1)$ al variare di z.

Per le applicazioni statistiche, è utile fornire i **valori critici** di una $N(0,1)$, cioè i valori z_α tali che $P(Z > z_\alpha) = \alpha$, con $\alpha \in (0, 0.5)$.

α	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
z_α	1.28	1.65	1.96	2.33	2.58	3.09	3.29

z_α individua una coda destra della distribuzione di probabilità di peso α, mentre $-z_\alpha$ individua una coda sinistra di peso α.

Esempio. *Pressione*. La variabile casuale X rappresenta la pressione sistolica, in mm di mercurio, di un generico individuo. Per la popolazione maschile italiana adulta si assume che $X \sim N(129, 392.04)$

Supponendo di scegliere a caso un individuo, si può calcolare

$$P(X < 135) = \Phi((135 - 129)/19.8) = \Phi(0.303) = 0.619.$$

$$P(120 < X < 130) = \Phi(0.051) - \Phi(-0.455) = 0.195.$$

La probabilità di selezionare un individuo con pressione compresa tra 120 e 150 mm, tra quelli con pressione superiore alla media, è

$$P(120 < X < 150 \mid X > 129) = \frac{P(120 < X < 150 \cap X > 129)}{P(X > 129)}$$
$$= \frac{P(129 < X < 150)}{1 - P(X \leq 129)} = \frac{\Phi(1.061) - \Phi(0)}{1 - \Phi(0)} = 0.711$$

$\Diamond$

Esempio. *Barre.* Per tagliare delle barre d'acciaio alla lunghezza nominale di 5 cm si utilizza un macchinario che fornisce barre con lunghezza $X \sim N(5.05, 0.01)$.

Ad un successivo controllo di qualità, si scartano le barre che differiscono dalla lunghezza nominale per più di un mm.

La probabilità che una generica barra soddisfi ai requisiti è

$$P(4.9 \leq X \leq 5.1) = \Phi(0.5) - \Phi(-1.5) = 0.625.$$

Se fosse possibile ricalibrare la procedura di taglio in modo da avere $\mu = 5$, si avrebbe

$$P(4.9 \leq X \leq 5.1) = \Phi(1) - \Phi(-1) = 0.683.$$

Infine, se si aumenta anche la precisione dello strumento di modo che $\sigma = 0.05$, si ha

$$P(4.9 \leq X \leq 5.1) = \Phi(2) - \Phi(-2) = 0.954.$$

$\Diamond$

Esempio. *Riso*. Un'industria alimentare confeziona pacchi di riso, con peso dichiarato pari a 500 gr, utilizzando un macchinario che fornisce pacchi con peso $X \sim N(500, 5)$.

Si scelgono a caso 5 confezioni di riso e si vuole calcolare la probabilità che non ci siano pacchi con peso inferiore a quello dichiarato.

La distribuzione di probabilità della variabile casuale Y, che conta il numero di pacchi di riso con peso inferiore a 500 gr, è $Bi(5, p)$, con $p = P(X \leq 500) = \Phi((500 - 500)/\sqrt{5}) = \Phi(0) = 0.5$.

Quindi la probabilità cercata è

$$P(Y = 0) = \left(\begin{array}{c} 5 \\ 0 \end{array} \right) 0.5^5 (1 - 0.5)^0 \doteq 0.03.$$

Inoltre, $E(Y) = 2.5$ e $V(Y) = 1.25$. $\Diamond$

Verifica di normalità

È utile effettuare un'analisi preliminare dei dati per verificare se il fenomeno in esame può essere descritto da un modello normale.

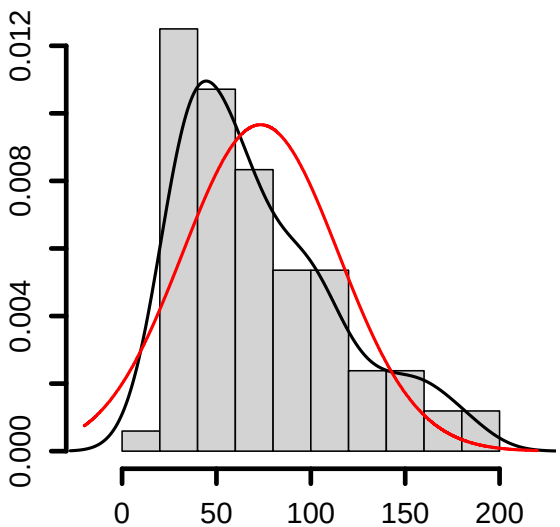
Si considerano i dati $x_1, \ldots, x_n$ e si vuole valutare se possono essere interpretati come osservazioni ripetute e indipendenti di una variabile casuale $X \sim N(\mu, \sigma^2)$.

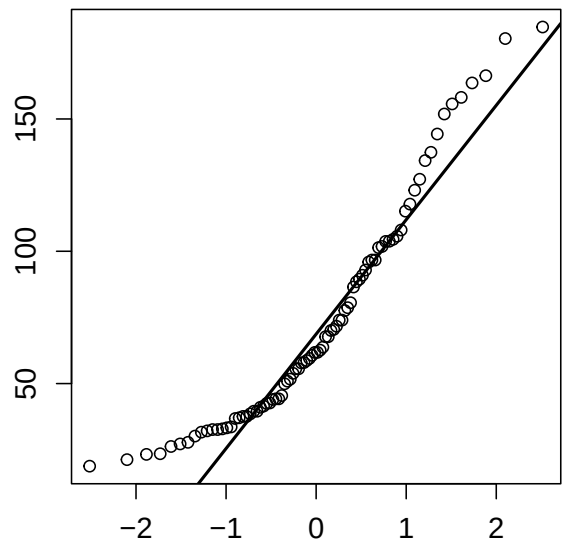
Tra le varie procedure per la **verifica di normalità** si presentano tre metodi grafici:

- confronto tra l'**istogramma** basato sui dati e la funzione di densità di una normale con media $\hat{\mu} = \sum_{i=1}^{n} x_i/n$ e varianza $\hat{\sigma}^2 = \sum_{i=1}^{n}(x_i - \hat{\mu})^2/n$ (media e varianza calcolate sui dati);
- confronto tra la **stima della densità** basata sui dati e la funzione di densità di una normale con media $\hat{\mu} = \sum_{i=1}^{n} x_i/n$ e varianza $\hat{\sigma}^2 = \sum_{i=1}^{n}(x_i - \hat{\mu})^2/n$
- rappresentazione dei quantili calcolati sui dati (in ordinata) e di quelli della distribuzione normale (in ascissa), detta **q-q plot**; se il modello normale è corretto i punti si trovano allineati su una linea retta.

Esempio. *Cipolle*. Si considerano $n = 84$ misurazioni sulla produzione (grammi per pianta) di una specie di cipolle bianche coltivate in Australia.

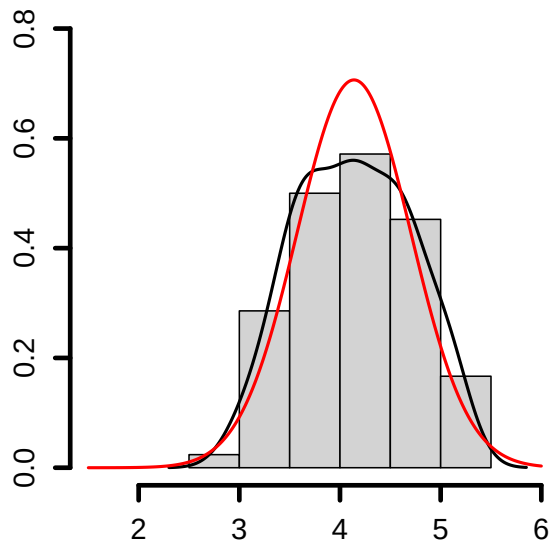
Dai dati si ricava che $\hat{\mu} = 73.33$ e $\hat{\sigma}^2 = 1704.28$. Si rappresenta l'istogramma delle frequenze relative e la stima della densità e si disegna sovrapposta la densità di una $N(\hat{\mu}, \hat{\sigma}^2)$ (in rosso). Inoltre, si confrontano quantili osservati e teorici utilizzando il q-q plot.

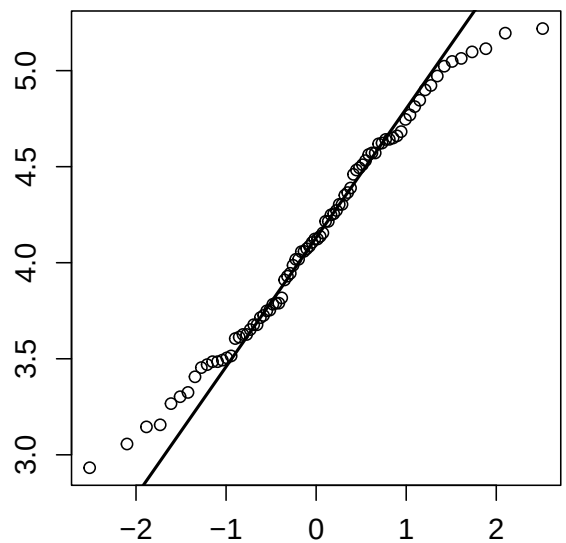




L'adattamento al modello normale non è soddisfacente. L'istogramma presenta una evidente asimmetria mentre il q-q plot evidenzia una notevole differenza sulle code della distribuzione.

Se si considera la trasformata logaritmica dei dati, $y_i = \log(x_i)$, $i = 1, \ldots, n$, si ottiene un adattamento un po' più soddisfacente alla distribuzione normale, come viene evidenziato nei grafici seguenti.





Soltanto sulle code la distribuzione normale non si adatta ancora bene ai dati. ◊

Modello chi-quadrato

Date le variabili casuali $Z_1, \ldots, Z_n$, $n \geq 1$, *indipendenti* (gli eventi associati a una generica variabile sono indipendenti dagli eventi associati alle altre) con *distribuzione* $N(0,1)$, allora la variabile casuale

$$Y = \sum_{i=1}^{n} Z_i^2$$

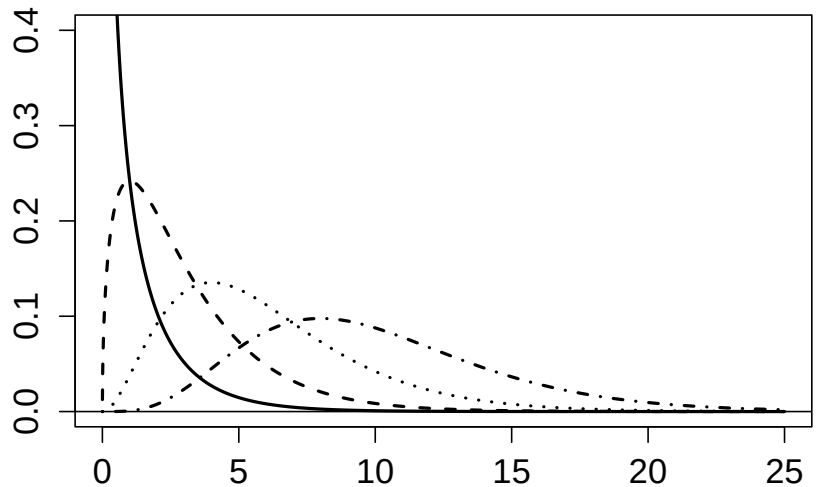
ha distribuzione **chi-quadrato** con n gradi di libertà, in simboli $Y \sim \chi^2(n)$.

È una variabile casuale continua con supporto è $S_Y = [0, +\infty)$ e $E(Y) = n$, $V(Y) = 2n$.

Se $Y_1 \sim \chi^2(n_1)$ e $Y_2 \sim \chi^2(n_2)$ sono indipendenti, la loro somma $Y_1 + Y_2 \sim \chi^2(n_1 + n_2)$.

Per $n \to +\infty$, la distribuzione di probabilità della variabile casuale $Y \sim \chi^2(n)$ tende alla distribuzione normale (l'approssimazione è buona per $n > 80$).

Grafico della funzione di densità della variabile casuale $Y \sim \chi^2(n)$ per $n = 1$ (—), $n = 3$ (– –), $n = 6$ ($\cdots$), $n = 10$ (- · -).



Utilizzando le **tavole statistiche** si ottengono i **valori critici** di una $Y \sim \chi^2(n)$, cioè i valori $\chi^2_{\alpha,n}$ tali che $P(Y > \chi^2_{\alpha,n}) = \alpha$, con $\alpha \in (0, 1)$, $n \geq 1$.

Modello *t* di Student

Date le variabili casuali $Z \sim N(0,1)$ e $Y \sim \chi^2(n)$ indipendenti, la variabile casuale

$$T = \frac{Z}{\sqrt{Y/n}}$$

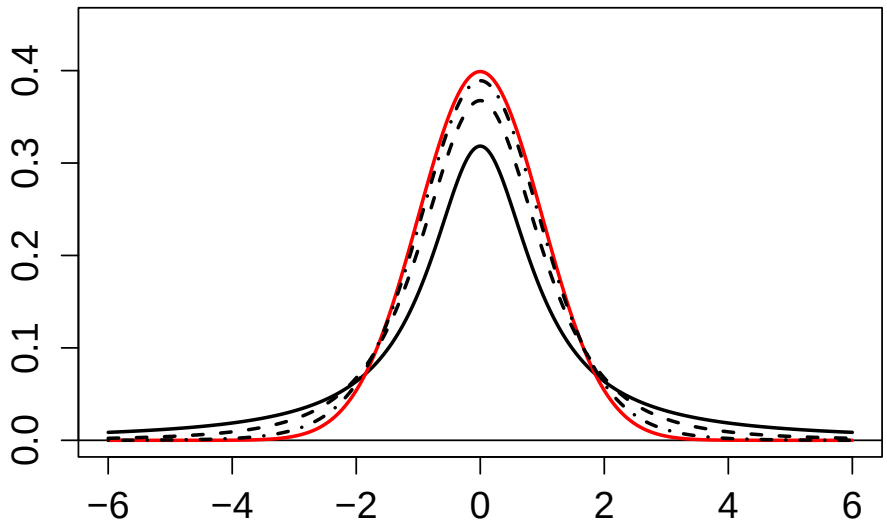
ha distribuzione ***t* di Student** con n gradi di libertà, in simboli $T \sim t(n)$.

È una variabile casuale continua con supporto è $S_T = \mathbf{R}$ e $E(T) = 0$, se $n > 1$, $V(T) = n/(n-2)$, se $n > 2$.

La funzione di densità è simmetrica rispetto alla retta $x = 0$ ed ha code più pesanti di quelle della normale standard.

Per $n \to +\infty$, la distribuzione di probabilità della variabile casuale $T \sim t(n)$ tende alla distribuzione normale (l'approssimazione è buona per $n > 30$).

Grafico della funzione di densità della variabile casuale $T \sim t(n)$ per $n = 1$ (—), $n = 3$ (– –), $n = 10$ (- · -) e della normale standard (in rosso).



Utilizzando le **tavole statistiche** si ottengono i **valori critici** di una $T \sim t(n)$, cioè i valori $t_{\alpha,n}$ tali che $P(T > t_{\alpha,n}) = \alpha$, con $\alpha \in (0, 0.5)$, $n \geq 1$. Per la simmetria, $t_{1-\alpha,n} = -t_{\alpha,n}$.

Modello F di Fisher

Date le variabili casuali $X \sim \chi^2(n)$ e $Y \sim \chi^2(m)$, $n, m \geq 1$, indipendenti, la variabile casuale

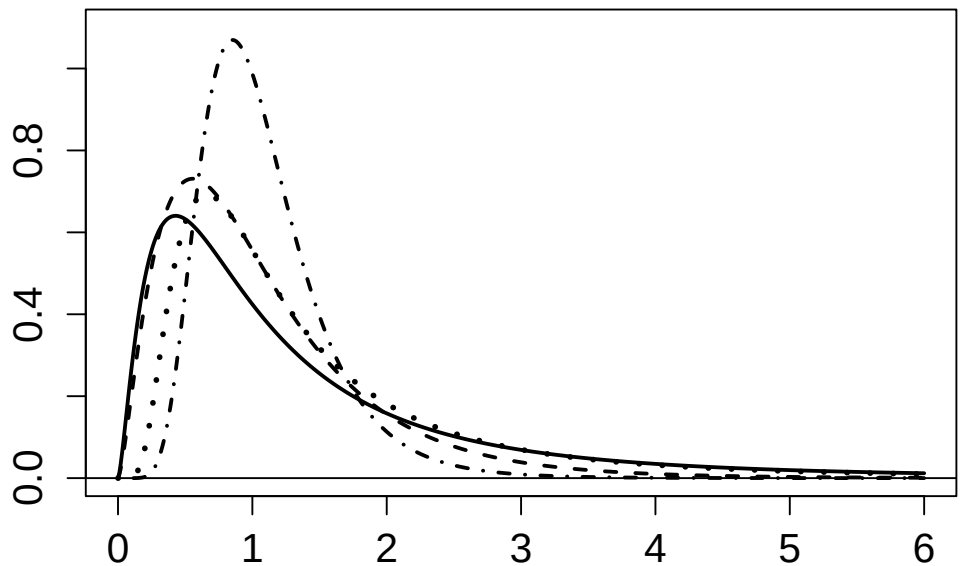
$$F = \frac{X/n}{Y/m}$$

ha distribuzione ***F* di Fisher** con n e m gradi di libertà, in simboli $F \sim F(n, m)$.

È una variabile casuale continua con supporto è $S_F = [0, +\infty)$ e $E(F) = m/(m-2)$, se $m > 2$.

Se $F \sim F(n, m)$, allora $F^{-1} \sim F(m, n)$. Se $T \sim t(n)$, allora $T^2 \sim F(1, n)$.

Grafico della funzione di densità della variabile casuale $F \sim F(n, m)$ per $n = 5$, $m = 5$ (—), $n = 5$, $m = 25$ (– –), $n = 25$, $m = 5$ $(\cdots)$, $n = 25$, $m = 25$ (- · -).



Utilizzando le **tavole statistiche** si ottengono i **valori critici** di una $F \sim F(n, m)$, cioè i valori $F_{\alpha,n,m}$ tali che $P(F > F_{\alpha,n,m}) = \alpha$, con $\alpha \in (0, 1)$, $n, m \geq 1$.