

STATISTICA E ANALISI DEI DATI (SECONDA PARTE)

Amelia Giuseppina Nobile¹

(a.a. 2018/2019)

30 novembre 2018

¹Dipartimento di Informatica, Università degli Studi di Salerno

Introduzione: Parte 2

L'indagine statistica è sempre effettuata su un insieme di entità (individui, oggetti,...) su cui si manifesta il fenomeno che si studia. Questo insieme è detto *popolazione* o *universo* e può essere costituito da un numero finito oppure infinito di unità; nel primo caso si parla di popolazione finita e nel secondo caso di popolazione illimitata. La conoscenza delle caratteristiche di una popolazione finita può essere ottenuta osservando la totalità delle entità della popolazione oppure un sottoinsieme di questa, detto *campione* estratto dalla popolazione. Una popolazione illimitata può invece essere studiata soltanto tramite un *campione* estratto dalla popolazione.

Di particolare importanza in statistica è l'*inferenza statistica*. Essa ha lo scopo di *estendere le misure ricavate dall'esame di un campione alla popolazione da cui il campione è stato estratto*.

Uno dei problemi centrali dell'inferenza statistica è il seguente: *si desidera studiare una popolazione descritta da una variabile aleatoria osservabile X la cui funzione di distribuzione ha una forma nota ma contiene un parametro $\vartheta \in \Theta$ non noto (o più parametri non noti)*.

Il termine *osservabile* significa che si possono osservare i valori assunti dalla variabile aleatoria X (ad esempio, eseguendo un esperimento casuale) e quindi il parametro non noto è presente soltanto nella legge di probabilità (funzione di distribuzione, funzione di probabilità, densità di probabilità). Ovviamente se ϑ è noto la legge di probabilità è completamente specificata.

Per ottenere informazioni sul parametro non noto ϑ della popolazione, si può fare uso dell'inferenza statistica considerando un campione estratto dalla popolazione e effettuando su tale campione delle opportune misure. Affinché le conclusioni dell'inferenza statistica siano valide il campione deve essere scelto in modo tale da essere *rappresentativo della popolazione*.

L'inferenza statistica si basa su due metodi fondamentali di indagine: la *stima dei parametri* e la *verifica delle ipotesi*.

La *stima dei parametri* ha lo scopo di determinare i valori non noti dei parametri di una popolazione (come il valore medio, la varianza,...) per mezzo dei corrispondenti parametri derivati dal campione estratto dalla popolazione (come la *media campionaria*, la *varianza campionaria*,...). Si possono usare *stime puntuali* o *stime per intervallo*.

Si parla di *stima puntuale* quando si stima un parametro non noto di una popolazione usando un singolo valore reale.

Alla stima puntuale di un parametro non noto di una popolazione (costituita da un unico valore) spesso si preferisce sostituire un intervallo di valori, detto *intervallo di confidenza*, ossia si cerca di determinare in base al campione osservato (x_1, x_2, \dots, x_n) due limiti (uno inferiore e uno superiore) entro i quali sia compreso il parametro non noto con un certo *grado di confidenza*, detto anche *grado di fiducia*.

La *verifica delle ipotesi* è un procedimento che consiste nel fare una *congettura* o un'ipotesi *sul parametro non noto ϑ o sulla distribuzione di probabilità* e nel decidere, sulla base del campione estratto se essa è accettabile. Spesso lo spazio Θ dei parametri, ossia l'insieme in cui può variare il parametro non noto della popolazione, si suddivide in due sottoinsiemi disgiunti Θ_0 e Θ_1 tali che $\Theta = \Theta_0 \cup \Theta_1$. L'ipotesi H_0 soggetta a verifica su ϑ consiste nell'affermare che $\vartheta \in \Theta_0$ ed è detta *ipotesi nulla*, mentre nell'ipotesi alternativa H_1 si assume invece che $\vartheta \in \Theta_1$. Il problema della verifica delle ipotesi consiste allora nel suddividere, mediante opportuni criteri, l'insieme dei possibili campioni in due sottoinsiemi, un sottoinsieme A di accettazione dell'ipotesi nulla e un sottoinsieme R di rifiuto dell'ipotesi nulla. Se il campione osservato $(x_1, x_2, \dots, x_n) \in A$ si accetta come valida l'affermazione che $\vartheta \in \Theta_0$, mentre se $(x_1, x_2, \dots, x_n) \in R$ si rifiuta l'ipotesi che $\vartheta \in \Theta_0$ e si accetta l'ipotesi alternativa che $\vartheta \in \Theta_1$.

Per affrontare i problemi di stima (puntuale o per intervallo) dei parametri e della verifica delle ipotesi statistiche nei prossimi due capitoli introdurremo le principali variabili aleatorie discrete e continue con l'ausilio di R. Successivamente ci occuperemo della stima puntuale e per intervallo e affronteremo alcuni problemi di verifica di ipotesi statistiche utilizzando R.

Capitolo 6

Variabili aleatorie discrete con R

6.1 Introduzione

Il sistema R mette a disposizione per ciascuna delle principali distribuzioni di probabilità discrete teoriche:

- la funzione di probabilità
- la funzione di distribuzione;
- la funzioni quantili;
- la funzione che simula tale variabile aleatoria mediante la generazione di numeri pseudocasuali;

Tutte queste funzioni utilizzano nomi che iniziano con una particolare lettera dell'alfabeto, in modo da indicare il tipo di funzione a cui fa riferimento, seguita dal nome della distribuzione teorica scelta. La particolare lettera dell'alfabeto può essere:

d calcola la funzione di probabilità di una variabile aleatoria in uno specifico punto o in un insieme di punti (*density mass*);

p calcola la funzione di distribuzione di una variabile aleatoria in uno specifico punto o in un insieme di punti (*probability distribution*);

q calcola la funzioni quantili;

r calcola la funzione che simula una variabile aleatoria mediante la generazione di numeri pseudocasuali.

In questo capitolo considereremo le seguenti distribuzioni discrete:

- distribuzione di Bernoulli;
- distribuzione binomiale;
- distribuzione geometrica e di Pascal;
- distribuzione ipergeometrica;
- distribuzione di Poisson.

6.2 Distribuzione di Bernoulli

Una prova di Bernoulli è un esperimento casuale caratterizzato da due soli possibili risultati, interpretabili l'uno come *successo* e l'altro come *insuccesso*, che si verificano rispettivamente con probabilità p e $1 - p$, con $0 < p < 1$. La variabile aleatoria X che descrive il risultato di una prova di Bernoulli assume soltanto due valori: 1 (indicante il successo) con probabilità p e 0 (indicante l'insuccesso) con probabilità $1 - p$.

Definizione 6.1 Una variabile aleatoria X di funzione di probabilità

$$p_X(x) = \begin{cases} 1 - p, & x = 0 \\ p, & x = 1 \\ 0, & \text{altrimenti,} \end{cases} \quad (6.1)$$

con $0 < p < 1$, è detta avere distribuzione di Bernoulli di parametro p .

La funzione di distribuzione di X è pertanto

$$F_X(x) = P(X \leq x) = \begin{cases} 0, & x < 0 \\ 1 - p, & 0 \leq x < 1 \\ 1, & x \geq 1. \end{cases} \quad (6.2)$$

Con la notazione $X \sim \mathcal{B}(1, p)$ intenderemo che X è una variabile aleatoria avente distribuzione di Bernoulli di parametro p , che chiameremo anche *variabile di Bernoulli*.

6.3 Distribuzione binomiale

Consideriamo l'esperimento consistente in n prove di Bernoulli indipendenti ed effettuate tutte in condizioni identiche, ed assumiamo che in ogni prova i risultati di interesse siano sintetizzabili nel verificarsi dei seguenti due eventi necessari ed incompatibili: A (interpretabile come successo) e \bar{A} (interpretabile come insuccesso), con $P(A) = p$ ($0 < p < 1$). Un siffatto esperimento si dice costituito da *n prove ripetute di Bernoulli*.

Sia X la variabile aleatoria che rappresenta il *numero di volte in cui si verifica l'evento A nelle n prove*.

Definizione 6.2 Una variabile aleatoria X di funzione di probabilità

$$p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, \dots, n \\ 0, & \text{altrimenti,} \end{cases} \quad (6.3)$$

con $0 < p < 1$ e n intero positivo, è detta avere distribuzione binomiale di parametri n e p .

Con la notazione $X \sim \mathcal{B}(n, p)$ intenderemo che X è una variabile aleatoria con distribuzione binomiale di parametri n e p , che chiameremo anche *variabile binomiale*. Nel caso particolare $n = 1$, la (6.3) si riduce alla funzione di probabilità di Bernoulli di parametro p . Dalla (6.3) si ricava:

$$\frac{p_X(r)}{p_X(r-1)} = \frac{p}{1-p} \frac{n-r+1}{r}, \quad r = 1, 2, \dots, n, \quad (6.4)$$

da cui segue che le probabilità binomiali (6.3) sono calcolabili ricorsivamente al seguente modo:

$$p_X(0) = (1-p)^n, \quad p_X(r) = \frac{p}{1-p} \frac{n-r+1}{r} p_X(r-1), \quad r = 1, 2, \dots, n.$$

La funzione di distribuzione di X è poi immediatamente ottenibile:

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i}, & k \leq x < k+1 \quad (k = 0, 1, \dots, n-1) \\ 1, & x \geq n. \end{cases} \quad (6.5)$$

Per una variabile aleatoria binomiale si ha:

$$E(X) = np, \quad \text{Var}(X) = np(1-p). \quad (6.6)$$

Se $p = 1/2$ la funzione di probabilità binomiale è simmetrica rispetto al suo valore medio $E(X) = n/2$.

R permette di calcolare la funzione di probabilità, la funzione di distribuzione e i quantili di una variabile aleatoria binomiale e anche di simulare tale variabile.

Si può richiedere ad R di eseguire direttamente il calcolo delle probabilità binomiali utilizzando la funzione

```
dbinom(x, size, prob)
```

Gli argomenti di tale funzione sono

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria binomiale considerata;
- $size$ è il numero complessivo delle prove;

- `prob` è la probabilità di successo in ciascuna prova.

Ad esempio, se $n = 5$ e $p = 0.95$ le probabilità binomiali possono essere così valutate:

```
> x<-0:5
> dbinom(x,size=5,prob=0.95)
[1] 0.0000003125 0.0000296875 0.0011281250 0.0214343750
[5] 0.2036265625 0.7737809375
```

ed è possibile arrotondare tali probabilità alla quarta cifra decimale nel seguente modo:

```
> x<-0:5
> round(dbinom(x,size=5,prob=0.95),4)
[1] 0.0000 0.0000 0.0011 0.0214 0.2036 0.7738
```

Le seguenti linee di codice permettono di visualizzare le funzioni di probabilità binomiali di Figura 6.1.

```
> par(mfrow=c(2,2))
> x<-0:5
> plot(x,dbinom(x,size=5,prob=0.95),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="n=5,p=0.95")
>
> x<-0:10
> plot(x,dbinom(x,size=10,prob=0.05),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="n=10,p=0.05")
>
> x<-0:20
> plot(x,dbinom(x,size=20,prob=0.5),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="n=20,p=0.5")
>
> x<-0:20
> plot(x,dbinom(x,size=20,prob=0.2),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="n=20,p=0.2")
```

Si può richiedere ad R di eseguire direttamente il calcolo della funzione di distribuzione binomiale utilizzando la funzione

```
pbinom(x, size, prob, lower.tail = TRUE)
```

Gli argomenti di tale funzione sono

- `x` è il valore assunto (o i valori assunti) dalla variabile aleatoria binomiale considerata;
- `size` è il numero complessivo delle prove;
- `prob` è la probabilità di successo in ciascuna prova;
- `lower.tail` se tale parametro è `TRUE` (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è `FALSE` calcola $P(X > x)$.

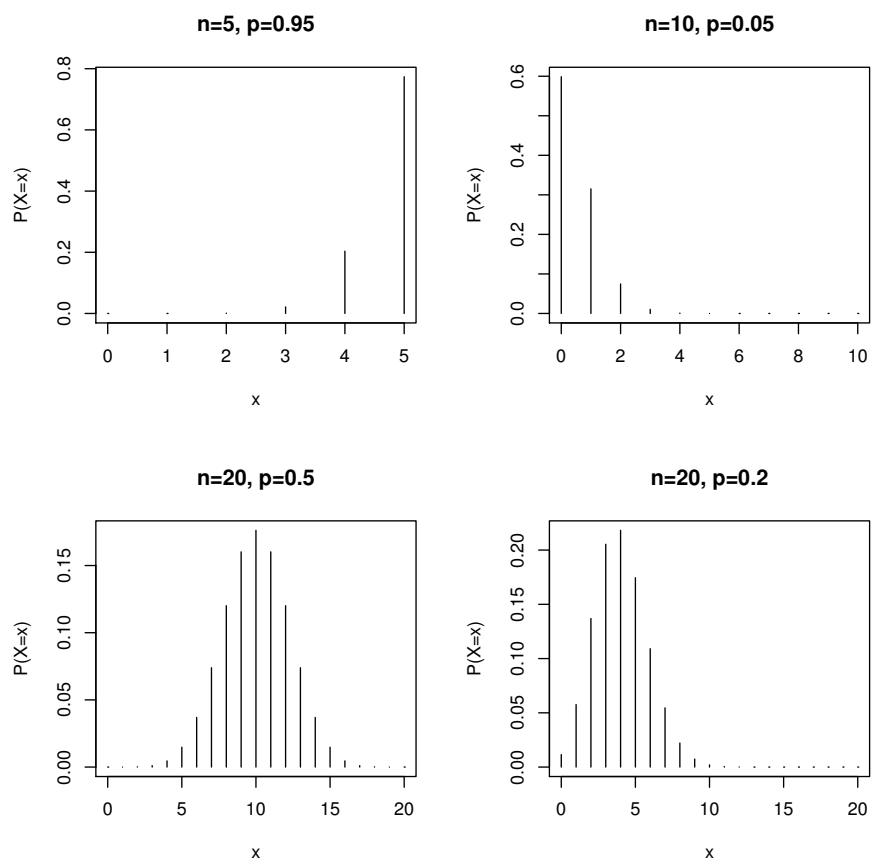


Figura 6.1: Rappresentazione della funzione di probabilità binomiale per alcuni valori di n e p .

232 CAPITOLO 6. VARIABILI ALEATORIE DISCRETE CON R

Ad esempio, se $n = 5$ e $p = 0.95$ la funzione di distribuzione binomiale può essere così valutata:

```
> x<-0:5
> pbinom(x,size=5,prob=0.95)
[1] 0.0000003125 0.0000300000 0.0011581250 0.0225925000
[5] 0.2262190625 1.0000000000
```

i cui risultati sono le probabilità

$$P(X \leq x) = \sum_{n=0}^x P(X = n), \quad x = 0, 1, \dots, 5.$$

Inoltre, se $n = 5$ e $p = 0.95$ le seguenti linee di codice

```
> x<-0:5
> pbinom(x,size=5,prob=0.95,lower.tail=FALSE)
[1] 0.9999997 0.9999700 0.9988419 0.9774075 0.7737809 0.0000000
```

mostrano le probabilità:

$$P(X > x) = 1 - P(X \leq x), \quad x = 0, 1, \dots, 5.$$

Le seguenti linee di codice permettono di visualizzare le funzioni di distribuzione di Figura 6.2.

```
> par(mfrow=c(2,2))
> x<-0:5
> plot(x,pbinom(x,size=5,prob=0.95),
+ xlab="x",ylab=expression(P(X<=x)),ylim=c(0,1),type="s",
+ main="n=5,p=0.95")
>
> x<-0:10
> plot(x,pbinom(x,size=10,prob=0.05),
+ xlab="x",ylab=expression(P(X<=x)),ylim=c(0,1),type="s",
+ main="n=10,p=0.05")
>
> x<-0:20
> plot(x,pbinom(x,size=20,prob=0.5),
+ xlab="x",ylab=expression(P(X<=x)),ylim=c(0,1),type="s",
+ main="n=20,p=0.5")
>
> x<-0:20
> plot(x,pbinom(x,size=20,prob=0.2),
+ xlab="x",ylab=expression(P(X<=x)),ylim=c(0,1),type="s",
+ main="n=20,p=0.2")
```

In R è possibile valutare il valore medio, la varianza, la deviazione standard e il coefficiente di variazione della distribuzione binomiale. Ad esempio, se $n = 20$ e $p = 0.2$ le seguenti linee di codice

```
> x<-0:20
> M1<-sum(x*dbinom(x,size=20,prob=0.2))
> M2<-sum(x^2*dbinom(x,size=20,prob=0.2))
> V<-M2-M1^2
> c(M1,V,sqrt(V),sqrt(V)/M1)
[1] 4.0000000 3.2000000 1.7888544 0.4472136
```

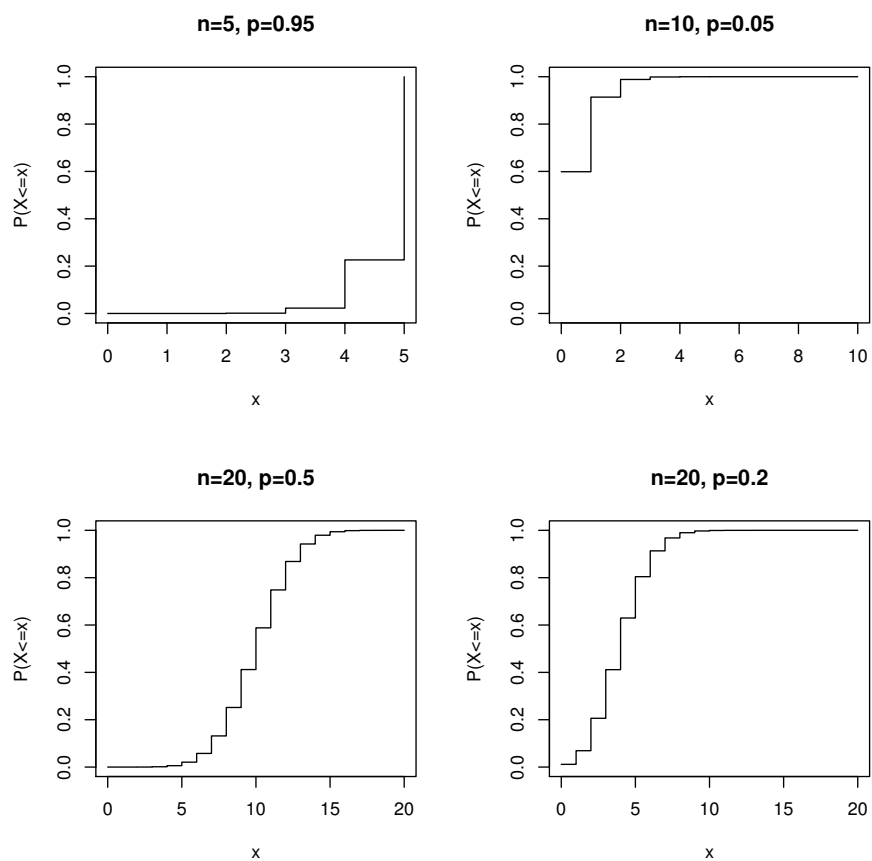


Figura 6.2: Rappresentazione della funzione di distribuzione binomiale per alcuni valori di n e p .

mostrano che $E(X) = 4$, $\text{Var}(X) = 3.2$, $\sqrt{\text{Var}(X)} = 1.7888544$ e $\text{CV}(X) = 0.4472136$. Si poteva, più semplicemente, utilizzare direttamente la (6.6) avendosi $E(X) = np = 20 \cdot 0.2 = 4$ e $\text{Var}(X) = np(1-p) = 20 \cdot 0.2 \cdot 0.8 = 3.2$.

In R si possono calcolare anche i quantili (percentili) della distribuzione binomiale attraverso la funzione

```
qbinom(z, size, prob, lower.tail = TRUE)
```

dove

- z è il valore assunto (o i valori assunti) dalle probabilità relative al percentile $z \cdot 100$ -esimo;
- $size$ è il numero complessivo delle prove;
- $prob$ è la probabilità di successo in ciascuna prova;
- `lower.tail` se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

Il risultato della funzione è il *percentile* $z \cdot 100$ -esimo, ossia il più piccolo numero intero k assunto dalla variabile aleatoria binomiale X tale che

$$P(X \leq k) \geq z \quad (k = 0, 1, \dots, n). \quad (6.7)$$

Ad esempio, se $n = 20$ e $p = 0.2$ le seguenti linee di codice forniscono i quartili Q_0, Q_1, Q_2, Q_3, Q_4

```
>z<-c(0,0.25,0.5,0.75,1)
> qbinom(z,size=20,prob=0.2)
[1] 0 3 4 5 20
```

che mostra che il primo quartile (25-esimo percentile) è $Q_1 = 3$, il secondo quartile o mediana (50-esimo percentile) è $Q_2 = 4$ e il terzo quartile (75-esimo percentile) è $Q_3 = 5$. Il minimo è $Q_0 = 0$ e il massimo è $Q_4 = 20$.

È possibile simulare in R la variabile aleatoria binomiale generando una sequenza di numeri pseudocasuali mediante la funzione

```
rbinom(N, size, prob)
```

dove

- N è lunghezza della sequenza da generare;
- $size$ è il numero complessivo delle prove;
- $prob$ è la probabilità di successo in ciascuna prova.

Ad esempio, se desideriamo simulare una variabile aleatoria binomiale $X \sim \mathcal{B}(20, 0.2)$ generando una sequenza di 50 numeri pseudocasuali (sequenza dei numeri di successi in 20 prove indipendenti di Bernoulli) si ha:

```

> sim<-rbinom(50,size=20,prob=0.2)
> sim
[1] 5 4 1 3 2 7 4 1 3 3 2 4 3 4 3 3 6 7 3 5 1 2 5 8 5 5 4 6 1 7 5 3
[33] 3 7 1 4 4 8 2 6 5 1 3 1 4 3 6 4 1 3
> table(sim)
sim
 1  2  3  4  5  6  7  8
 8  4 12  9  7  4  4  2
> table(sim)/length(sim)
sim
 1    2    3    4    5    6    7    8
0.16 0.08 0.24 0.18 0.14 0.08 0.08 0.04

```

dove `table(sim)/length(sim)` fornisce le frequenze relative con cui i numeri $0, 1, \dots, 20$ si presentano nella sequenza generata. Occorre sottolineare che differenti esecuzioni conducono a sequenze pseudocasuali diverse.

Il codice seguente permette di confrontare la funzione di probabilità binomiale teorica di una variabile binomiale $X \sim \mathcal{B}(20, 0.2)$ con quella simulata all'aumentare della lunghezza $N = 500, 5000, 50000$ della sequenza generata.

```

> par(mfrow=c(2,2))
> x<-0:20
> plot(x,dbinom(x,size=20,prob=0.2),
+ xlab="x",type="h",ylab="Probabilità",
+ main="n=20,p=0.2",ylim=c(0,0.24))
>
> sim1<-rbinom(500,size=20,prob=0.2)
> plot(table(sim1)/length(sim1),
+ xlab="x",type="h",ylab="Frequenza relativa",xlim=c(0,20),
+ main="n=20,p=0.2,N=500",ylim=c(0,0.24))
>
> sim2<-rbinom(5000,size=20,prob=0.2)
> plot(table(sim2)/length(sim2),
+ xlab="x",type="h",ylab="Frequenza relativa",xlim=c(0,20),
+ main="n=20,p=0.2,N=5000",ylim=c(0,0.24))
>
> sim3<-rbinom(50000,size=20,prob=0.2)
> plot(table(sim3)/length(sim3),
+ xlab="x",type="h",ylab="Frequenza relativa",xlim=c(0,20),
+ main="n=20,p=0.2,N=50000",ylim=c(0,0.24))

```

Si nota che all'aumentare della lunghezza della sequenza generata il grafico delle frequenze relative si avvicina sempre di più al grafico della funzione di probabilità binomiale.

Esempio 6.1 (*Regola di decisione a maggioranza*) Supponiamo di effettuare n lanci ($n = 3, 5, \dots$) indipendenti di una moneta e supponiamo che sia $p = 0.7$ la probabilità di successo e $1 - p = 0.3$ la probabilità di insuccesso in ogni singola prova. Sia X la variabile aleatoria che descrive il numero di successi in ogni singola prova. Desideriamo calcolare la probabilità Q_n che il numero di successi sia maggiore del numero di insuccessi nelle n prove. Ciò si verifica se $X > n - X$, ossia $X > n/2$ o equivalentemente $X \geq (n+1)/2$. Essendo le prove

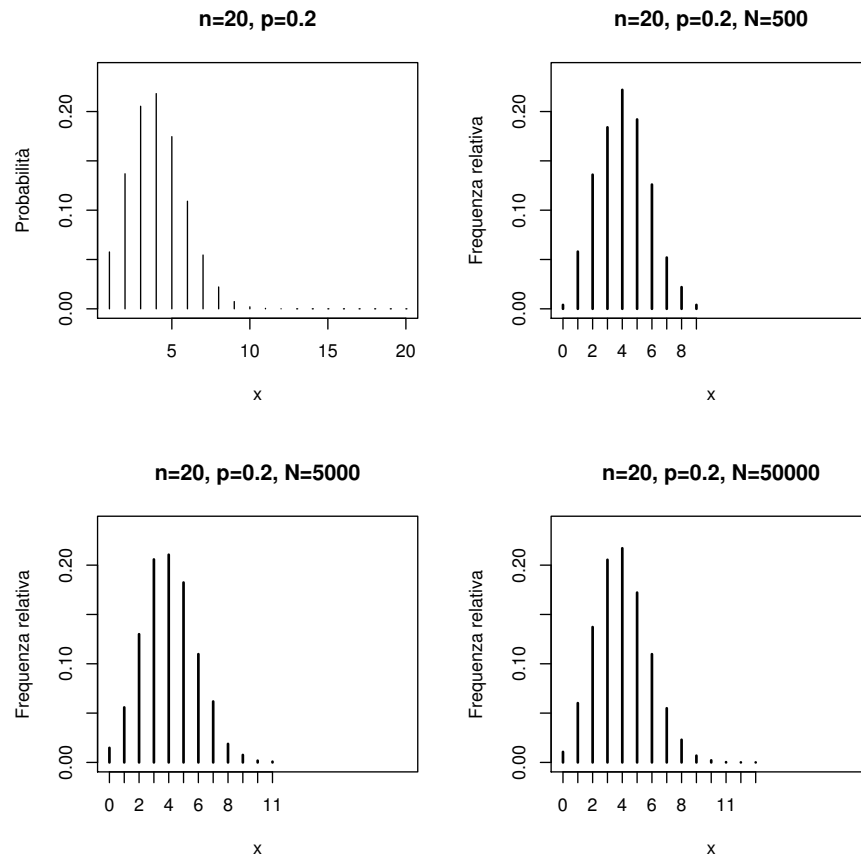


Figura 6.3: Confronto della funzione di probabilità binomiale teorica e delle frequenze relative simulate per una variabile aleatoria binomiale $X \sim \mathcal{B}(20, 0.2)$.

indipendenti, X ha distribuzione binomiale di parametri n e $p = 0.7$:

$$Q_n = P\left(X \geq \frac{n+1}{2}\right) = \sum_{k=(n+1)/2}^n \binom{n}{k} (0.7)^k (0.3)^{n-k} \quad (n = 3, 5, \dots),$$

o equivalentemente

$$\begin{aligned} Q_n &= P\left(X \geq \frac{n+1}{2}\right) = 1 - P\left(X < \frac{n+1}{2}\right) = 1 - P\left(X \leq \frac{n+1}{2} - 1\right) \\ &= 1 - F_X\left(\frac{n+1}{2} - 1\right) \end{aligned}$$

In R il codice per calcolare la probabilità di una corretta classificazione quando il numero di prove è $3, 5, \dots, 35$ è

```
> n<-seq(3,35,2)
> med<-(n+1)/2
> Qn<-1-pbinom(med-1,n,0.7)
> round(Qn,4)
[1] 0.7840 0.8369 0.8740 0.9012 0.9218 0.9376 0.9500 0.9597 0.9674
[10] 0.9736 0.9786 0.9825 0.9857 0.9883 0.9905 0.9922 0.9936
>
> plot(n,Qn,xlab="Numero di prove",
+ ylab=expression('Q'['n']),
+ ylim=c(0.7,1),type="h")
```

e tali probabilità sono rappresentate in Figura 6.4. Si nota che la probabilità Q_n aumenta con il numero di prove e presenta una rapida crescita fino a circa 21 prove, dopo di che la crescita delle probabilità diventa molto più lenta.

6.4 Distribuzione geometrica e di Pascal

Descriviamo la distribuzione geometrica e quella di Pascal, mostrandone le differenze.

⇒ **Distribuzione geometrica**

Si consideri l'esperimento consistente in una successione di prove ripetute di Bernoulli di parametro $p \in (0, 1)$. Si supponga di essere interessati all'evento

$$E_r = \{\text{il primo successo si verifica alla prova } r\text{-esima}\} \quad (r = 1, 2, \dots).$$

Dall'ipotesi di indipendenza delle prove si ricava che $P(E_r) = (1-p)^{r-1}p$. Sia X la variabile aleatoria che descrive il numero di prove necessarie per ottenere il primo successo; è evidente che $P(X=r) = P(E_r)$ per $r = 1, 2, \dots$.

Definizione 6.3 Una variabile aleatoria X di funzione di probabilità

$$p_X(x) = \begin{cases} p(1-p)^{x-1}, & x = 1, 2, \dots \\ 0, & \text{altrimenti,} \end{cases} \quad (6.8)$$

con $0 < p < 1$ si dice avere distribuzione geometrica di parametro p .

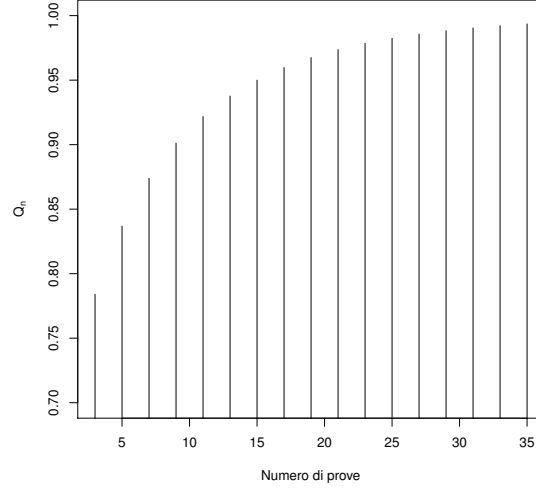


Figura 6.4: Probabilità Q_n al variare del numero di prove $n = 3, 5, \dots, 35$, assumendo che $p = 0.7$, utilizzando una regola di decisione a maggioranza.

Dalla (6.8) segue immediatamente che $p_X(r)$ è strettamente decrescente in $r = 1, 2, \dots$. Poiché

$$\sum_{r=1}^k p_X(r) = \sum_{r=1}^k p(1-p)^{r-1} = p \sum_{s=0}^{k-1} (1-p)^s = p \frac{1 - (1-p)^k}{1 - (1-p)} = 1 - (1-p)^k,$$

la funzione di distribuzione di X è la seguente:

$$F_X(x) = \begin{cases} 0, & x < 1 \\ 1 - (1-p)^k, & k \leq x < k+1 \end{cases} \quad (k = 1, 2, \dots). \quad (6.9)$$

Per una variabile aleatoria geometrica si ha:

$$E(X) = \frac{1}{p}, \quad \text{Var}(X) = \frac{1-p}{p^2}. \quad (6.10)$$

Un'interessante proprietà della distribuzione geometrica è la seguente.

Proposizione 6.1 *Se X è una variabile aleatoria con distribuzione geometrica, risulta:*

$$P(X > r+n \mid X > r) = P(X > n), \quad (6.11)$$

con r e n interi non negativi.

Per illustrare il significato della (6.11) si consideri una successione di prove ripetute di Bernoulli. Se nelle prime r prove non si è avuto nessun successo, la

probabilità che non si verifichi alcun successo fino alla prova $r + n$ non dipende da r , ossia da quanto si è atteso, ma solo dal numero n di prove ancora da effettuarsi. La (6.11) esprime dunque la proprietà di *assenza di memoria* della distribuzione geometrica.

⇒ **Distribuzione di Pascal**

Una *variabile aleatoria di Pascal* è invece definita come $Y = X - 1$; essa assume valori $0, 1, \dots$ e rappresenta il *numero di fallimenti che si verificano prima del primo successo*. Per la variabile aleatoria di Pascal Y si ha:

$$p_Y(y) = P(Y = y) = P(X = y + 1) = p_X(y + 1) = \begin{cases} p(1-p)^y & y = 0, 1, \dots \\ 0, & \text{altrimenti,} \end{cases} \quad (6.12)$$

Poiché

$$\sum_{r=0}^k p_Y(y) = \sum_{r=0}^k p(1-p)^r = 1 - (1-p)^{k+1},$$

la funzione di distribuzione della variabile aleatoria di Pascal Y è la seguente:

$$F_Y(y) = P(Y \leq y) = \begin{cases} 0, & y < 0 \\ 1 - (1-p)^{k+1}, & k \leq y < k+1 \end{cases} \quad (k = 0, 1, \dots). \quad (6.13)$$

Dalla (6.12) segue immediatamente che $p_Y(r)$ è strettamente decrescente in $r = 0, 1, \dots$. Per una variabile aleatoria di Pascal Y si ha:

$$E(Y) = E(X - 1) = \frac{1-p}{p}, \quad \text{Var}(Y) = \text{Var}(X) = \frac{1-p}{p^2}. \quad (6.14)$$

In R sia per la variabile aleatoria di Pascal che per la variabile aleatoria geometrica si utilizzano le funzioni `dgeom()`, `pgeom()`, `qgeom()` e `rgeom()` come illustrato nel seguito.

6.4.1 Distribuzione di Pascal in R

R permette di calcolare la funzione di probabilità, la funzione di distribuzione e i quantili di una variabile aleatoria di Pascal Y e anche di simulare tale variabile. Si può richiedere ad R di eseguire direttamente il calcolo delle probabilità di una variabile aleatoria di Pascal Y utilizzando la funzione

`dgeom(x, prob)`

Gli argomenti di tale funzione sono

- `x` è il valore assunto (o i valori assunti) dalla variabile aleatoria di Pascal considerata;
- `prob` è la probabilità di successo in ciascuna prova.

Ad esempio, se $p = 0.95$ le probabilità di una variabile aleatoria di Pascal Y possono essere così valutate:

A.G. Nobile

```
> x<-0:5
> dgeom(x,prob=0.95)
[1] 9.50000e-01 4.75000e-02 2.37500e-03 1.18750e-04 5.93750e-06
[6] 2.96875e-07
```

ed è possibile arrotondare tali probabilità alla quarta cifra decimale nel seguente modo:

```
> x<-0:5
> round(dgeom(x,prob=0.95),4)
[1] 0.9500 0.0475 0.0024 0.0001 0.0000 0.0000
```

Le seguenti linee di codice permettono di visualizzare le funzioni di probabilità di una variabile aleatoria di Pascal Y come illustrato in Figura 6.5.

```
> par(mfrow=c(2,2))
> y<-0:5
> plot(y,dgeom(y,prob=0.95),
+ xlab="y",ylab="P(Y=y)",type="h",
+ main="p=0.95")
>
> y<-0:10
> plot(y,dgeom(y,prob=0.05),
+ xlab="y",ylab="P(Y=y)",type="h",
+ main="p=0.05")
>
> y<-0:20
> plot(y,dgeom(y,prob=0.5),
+ xlab="y",ylab="P(Y=y)",type="h",
+ main="p=0.5")
>
> y<-0:20
> plot(y,dgeom(y,prob=0.2),
+ xlab="y",ylab="P(Y=y)",type="h",
+ main="p=0.2")
```

Si può richiedere ad R di eseguire direttamente il calcolo della funzione di distribuzione di una variabile aleatoria di Pascal Y utilizzando la funzione

```
pgeom(x, prob, lower.tail = TRUE)
```

Gli argomenti di tale funzione sono

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria di Pascal considerata;
- $prob$ è la probabilità di successo in ciascuna prova;
- $lower.tail$ se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

Ad esempio, se $n = 5$ e $p = 0.95$ la funzione di distribuzione di una variabile aleatoria di Pascal può essere così valutata:

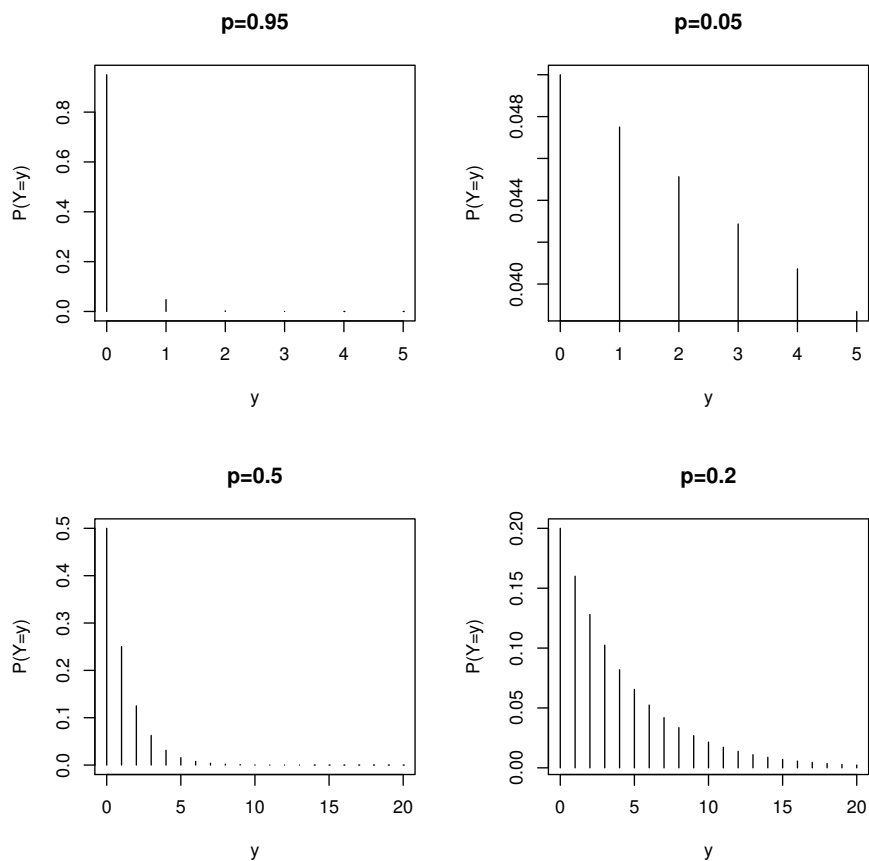


Figura 6.5: Rappresentazione della funzione di probabilità di Pascal Y per alcuni valori di p .

```
> x<-0:5
> pgeom(x,prob=0.95)
[1] 0.9500000 0.9975000 0.9998750 0.9999938 0.9999997 1.0000000
```

i cui risultati sono le probabilità

$$P(Y \leq x) = \sum_{n=0}^x P(Y = n), \quad x = 0, 1, \dots, 5.$$

Le seguenti linee di codice permettono di visualizzare le funzioni di distribuzione delle variabili aleatorie di Pascal di Figura 6.6.

```
> par(mfrow=c(2,2))
> y<-0:5
> plot(y,pgeom(y,prob=0.95),
+ xlab="y",ylab=expression(P(Y<=y)),ylim=c(0,1),type="s",
+ main="p=0.95")
>
> y<-0:50
> plot(y,pgeom(y,prob=0.05),
+ xlab="y",ylab=expression(P(Y<=y)),ylim=c(0,1),type="s",
+ main="p=0.05")
>
> y<-0:20
> plot(y,pgeom(y,prob=0.5),
+ xlab="y",ylab=expression(P(Y<=y)),ylim=c(0,1),type="s",
+ main="p=0.5")
>
> y<-0:20
> plot(y,pgeom(y,prob=0.2),
+ xlab="y",ylab=expression(P(Y<=y)),ylim=c(0,1),type="s",
+ main="p=0.2")
```

È possibile calcolare il valore medio, la varianza, la deviazione standard e il coefficiente di variazione della distribuzione di Pascal attraverso la (6.14). Ad esempio, se $p = 0.2$ si ricava $E(Y) = (1 - p)/p = 4.0$, $\text{Var}(Y) = (1 - p)/p^2 = 20$ e $\text{CV}(Y) = \sqrt{(20)/4} = 1.118034$.

In R si possono calcolare anche i quantili (percentili) della distribuzione di Pascal attraverso la funzione

```
qgeom(z, prob, lower.tail = TRUE)
```

dove

- z è il valore assunto (o i valori assunti) dalle probabilità relative al percentile $z \cdot 100$ -esimo;
- prob è la probabilità di successo in ciascuna prova;
- lower.tail se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

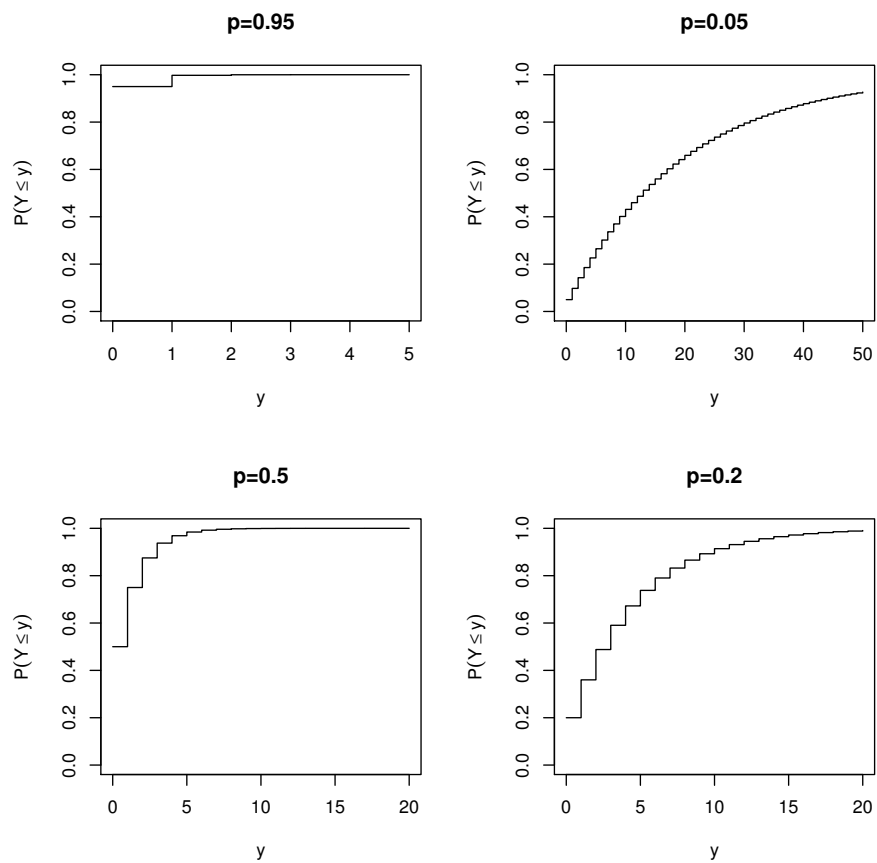


Figura 6.6: Rappresentazione della funzione di distribuzione di Pascal per alcuni valori di p .

Per una distribuzione di Pascal il percentile (quantile) $z \cdot 100$ -esimo è il più piccolo intero k tale che

$$P(Y \leq k) = 1 - (1 - p)^{k+1} \geq z \quad (k = 0, 1, \dots). \quad (6.15)$$

da cui segue

$$(1 - p)^{k+1} \leq 1 - z \quad (k = 0, 1, \dots),$$

Pertanto, per una distribuzione di Pascal il percentile (quantile) $z \cdot 100$ -esimo è il più piccolo intero k tale che

$$k \geq \frac{\ln(1 - z)}{\ln(1 - p)} - 1 \quad (k = 0, 1, \dots).$$

Se, ad esempio, si sceglie $p = 0.2$ si ha:

$$\begin{aligned} z = 0 &\implies k \geq \frac{\ln 1}{\ln 0.8} - 1 = -1 \implies Q_0 = 0, \\ z = 0.25 &\implies k \geq \frac{\ln 0.75}{\ln 0.8} - 1 = 0.2892 \implies Q_1 = 1, \\ z = 0.5 &\implies k \geq \frac{\ln 0.5}{\ln 0.8} - 1 = 2.1063 \implies Q_2 = 3, \\ z = 0.75 &\implies k \geq \frac{\ln 0.25}{\ln 0.8} - 1 = 5.2126 \implies Q_3 = 6, \\ z = 1 &\implies k \geq +\infty \implies Q_4 = +\infty, \end{aligned}$$

In R il risultato della funzione `qgeom()` è il percentile $z \cdot 100$ -esimo della distribuzione di Pascal. Ad esempio, se $p = 0.2$ le seguenti linee di codice forniscono i quartili Q_0, Q_1, Q_2, Q_3, Q_4

```
>z<-c(0,0.25,0.5,0.75,1)
> qgeom(z,prob=0.2)
[1] 0 1 3 6 Inf
```

che mostra che il primo quartile (il 25-esimo percentile) è $Q_1 = 1$, il secondo quartile o mediana (il 50-esimo percentile) è $Q_2 = 3$ e il terzo quartile (75-esimo percentile) è $Q_3 = 6$. Il minimo è $Q_0 = 0$ e il massimo è $Q_4 = \infty$.

È possibile simulare in R la variabile aleatoria di Pascal generando una sequenza di numeri pseudocasuali mediante la funzione

```
rgeom(N, prob)
```

dove

- `N` è lunghezza della sequenza da generare;
- `prob` è la probabilità di successo in ciascuna prova.

Ad esempio, se desideriamo generare una sequenza di 20 numeri pseudocasuali simulando una variabile aleatoria di Pascal con $p = 0.2$ si ha:

```

> sim<-rgeom(20,prob=0.2)
> sim
[1] 1 1 1 0 12 0 2 0 1 0 6 5 7 0 1 4 1 2 4 1
> table(sim)
sim
 0  1  2  4  5  6  7 12
 5  7  2  2  1  1  1  1
> table(sim)/length(sim)
sim
 0    1    2    4    5    6    7   12
0.25 0.35 0.10 0.10 0.05 0.05 0.05 0.05

```

dove `table(sim)/length(sim)` fornisce le frequenze relative con cui i numeri 0, 1, ... si presentano nella sequenza generata. Differenti esecuzioni possono condurre a diverse sequenze pseudocasuali.

6.4.2 Distribuzione geometrica in R

Per ottenere le funzioni di probabilità di una variabile aleatoria geometrica $X = Y + 1$ con le stesse scelte delle probabilità di Figura 6.5 occorre procedere come segue (vedi Figura 6.7)

```

> par(mfrow=c(2,2))
> x<-0:5
> plot(x+1,dgeom(x,prob=0.95),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="p=0.95")
>
> x<-0:10
> plot(x+1,dgeom(x,prob=0.05),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="p=0.05")
>
> x<-0:20
> plot(x+1,dgeom(x,prob=0.5),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="p=0.5")
>
> y<-0:20
> plot(x+1,dgeom(x,prob=0.2),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="p=0.2")

```

Possiamo procedere allo stesso modo per calcolare la funzione di distribuzione. Per i quantili di una variabile aleatoria geometrica X , basta aggiungere 1 ai quantili della variabile di Pascal.

Se invece si desidera simulare una variabile aleatoria geometrica X basta semplicemente ricordare che $X = Y + 1$, dove Y è una variabile aleatoria di Pascal. Ad esempio, se desideriamo generare una sequenza di 20 numeri pseudocasuali simulando una variabile aleatoria geometrica con $p = 0.2$ si ha:

```

> sim<-rgeom(20,prob=0.2)+1
> sim

```

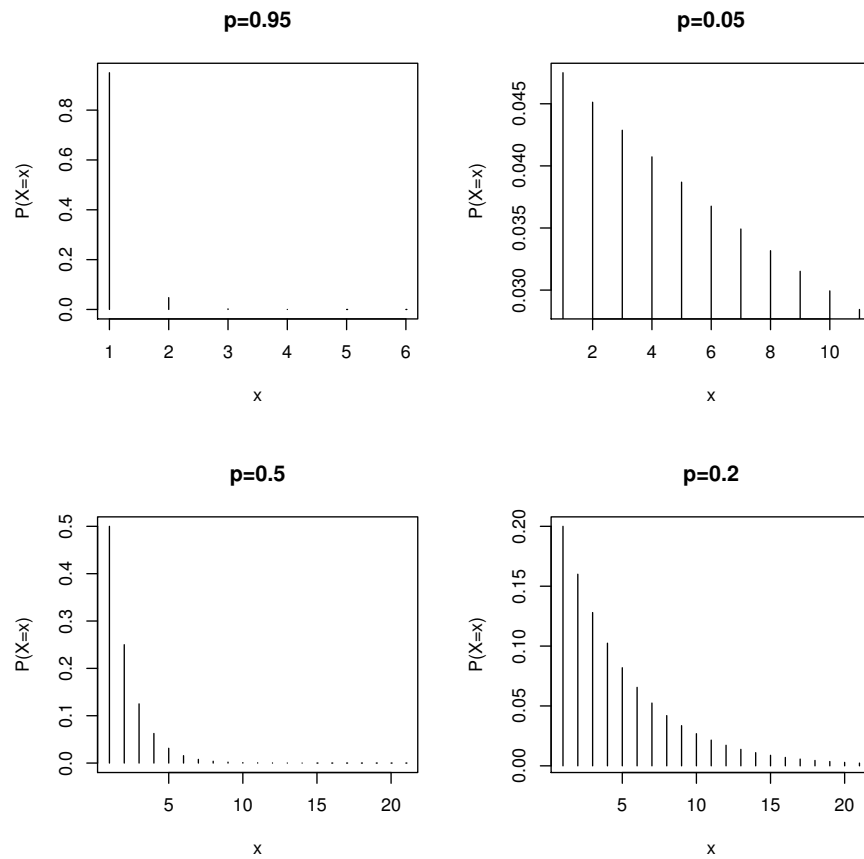


Figura 6.7: Rappresentazione della funzione di probabilità geometrica per alcuni valori di p .


```
[1] 28 1 2 1 9 2 2 1 5 12 9 1 2 7 2 4 3 1 21 2
> table(sim)
sim
 1  2  3  4  5  7  9 12 21 28
 5  6  1  1  1  1  2  1  1  1
> table(sim)/length(sim)
sim
 1      2      3      4      5      7      9      12      21      28
0.25 0.30 0.05 0.05 0.05 0.05 0.10 0.05 0.05 0.05
```

dove `table(sim)/length(sim)` fornisce le frequenze relative con cui i numeri $0, 1, \dots, 20$ si presentano nella sequenza generata. Si nota che differenti esecuzioni possono condurre a sequenze pseudocasuali diverse.

Il codice seguente permette di confrontare la funzione di probabilità teorica geometrica con quella simulata all'aumentare della lunghezza $N = 500, 5000, 50000$ della sequenza generata.

```
> par(mfrow=c(2,2))
> x<-0:20
> plot(x+1,dgeom(x,prob=0.2),xlab="x",ylab="Probabilita' ",
+ type="h",main="p=0.2",xlim=c(0,20))
>
> sim1<-rgeom(500,prob=0.2)+1
> plot(table(sim1)/length(sim1),xlab="x",type="h",
+ ylab="Frequenza relativa",xlim=c(0,20),ylim=c(0,0.20),
+ main="p=0.2, N=500")
>
> sim2<-rgeom(5000,prob=0.2)+1
> plot(table(sim2)/length(sim2),xlab="x",type="h",
+ ylab="Frequenza relativa",xlim=c(0,20),ylim=c(0,0.20),
+ main="p=0.2, N=5000")
>
> sim3<-rgeom(50000,prob=0.2)+1
> plot(table(sim3)/length(sim3),xlab="x",type="h",
+ ylab="Frequenza relativa",xlim=c(0,20),ylim=c(0,0.20),
+ main="p=0.2, N=50000")
```

Si nota che nel codice per rappresentare la funzione di probabilità geometrica si considerano punti di ascissa $x + 1$, (con $x = 0, 1, \dots$) a cui corrispondono ordinate $p(1 - p)^x$ ottenute tramite la distribuzione di Pascal.

Si nota che all'aumentare della lunghezza della sequenza generata il grafico delle frequenze relative si avvicina sempre di più al grafico della funzione di probabilità geometrica. Anche in questo caso possiamo confrontare le medie campionarie e le varianze campionarie delle sequenze geometriche simulate con la media $E(X) = 1/p$ e la varianza $Var(X) = (1 - p)/p^2$.

Esempio 6.2 Paradosso di San Pietroburgo (Daniele Bernoulli, 1700–1782)

Si consideri il gioco d'azzardo consistente in una successione di lanci indipendenti di una moneta (truccata o non truccata). Un giocatore viene ammesso al gioco previo pagamento di una certa somma, diciamo di s Euro. Si suppone che il giocatore riceve 2 Euro se si verifica testa al primo lancio, 4 Euro se testa si verifica per la prima volta al secondo lancio, 8 Euro se testa si verifica per la prima volta al terzo lancio e, in generale, 2^n Euro se testa si verifica per la prima

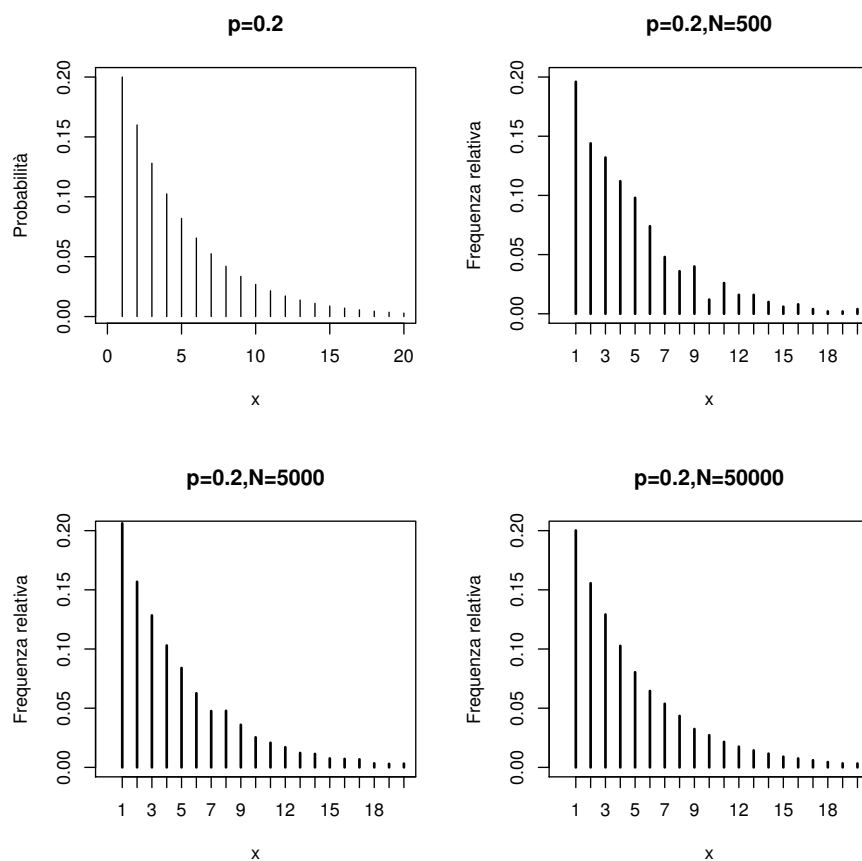


Figura 6.8: Confronto della funzione di probabilità geometrica teorica e delle frequenze relative simulate per una variabile aleatoria geometrica di parametro $p = 0.2$.

volta all' n -esimo lancio. Ci si chiede quale sia un valore “equo” di s , ossia quale sia un’equa somma da richiedersi al giocatore per consentirgli di partecipare al gioco. Intuitivamente si sarebbe portati ad identificare s con la somma che in media il giocatore vince. Denotando con X il guadagno del giocatore, risulta

$$E(X) = \sum_{n=1}^{+\infty} 2^n (1-p)^{n-1} p = \frac{p}{1-p} \sum_{n=1}^{+\infty} [2(1-p)]^n.$$

Si nota che tale somma converge se e solo se $2(1-p) < 1$, ossia se $p > 1/2$, e risulta

$$E(X) = \frac{p}{1-p} \left[\frac{1}{1-2(1-p)} - 1 \right] = \frac{p}{1-p} \frac{2(1-p)}{2p-1} = \frac{2p}{2p-1} \quad (p > 1/2).$$

Se $p \leq 1/2$, il guadagno medio del giocatore vale $+\infty$. Questo ultimo risultato è paradossale in quanto si esigerebbe una somma infinitamente grande per consentire la partecipazione ad un gioco dal quale non può che ricavarsi una vincita limitata.

Simuliamo ora il gioco valutando il guadagno medio. Sia (x_1, x_2, \dots, x_N) un campione di lunghezza N estratto da una popolazione geometrica, dove x_i denota il lancio i -esimo in cui è avvenuto il primo successo ($x_i = 1, 2, \dots$). Tale campione può essere ottenuto tramite la simulazione di una variabile aleatoria geometrica. Il guadagno medio ottenuto dal giocatore è allora

$$\bar{x} = \frac{2^{x_1} + 2^{x_2} + \dots + 2^{x_N}}{N}.$$

Ad esempio, effettuando $N = 100000$ simulazioni per $p = 0.5$ il codice seguente fornisce:

```
> ex<-rgeom(100000,prob=0.5)+1
> vinc<-2**ex
> mean(vinc)
[1] 15.72892
```

che mostra che la media simulata è finita mentre la media teorica diverge. Il paradosso, come molti di quelli che riguardano l’infinito, del tutto corretto in teoria, ma non funziona in pratica. Perché funzioni, infatti, richiede non solo la possibilità di proseguire il gioco per un tempo indefinito ma anche e soprattutto l’assunzione che il banco abbia a disposizione una riserva infinita di denaro. L’intuizione trova conferma in una simulazione a partire da 100000 fino a 200000 giocate per $p = 0.4, 0.5, 0.6, 0.9$ il cui codice è illustrato nel seguito:

```
>par(mfrow=c(2,2))
>n<-seq(100000,200000,1000)
>
>guad1<-function(k){
+mean(2**(rgeom(k,prob=0.4)+1))}
>gguad1<-sapply(seq(100000,200000,1000),guad1)
>plot(n,gguad1,xlab="Numero di simulazioni",
+ylab="Guadagno medio simulato",main="Paradosso di San Pietroburgo",
```

```

+p=0.4",type="l")
>
>guad2<-function(k){
+mean(2**(rgeom(k,prob=0.5)+1))}
>gguad2<-sapply(seq(100000,200000,1000),guad2)
>plot(n,gguad2,xlab="Numero di simulazioni",
+ylab="Guadagno medio simulato",main="Paradosso di San Pietroburgo",
+p=0.5",type="l")
>guad3<-function(k){
+mean(2**(rgeom(k,prob=0.6)+1))}
>
>guad3<-function(k){
+mean(2**(rgeom(k,prob=0.6)+1))}
>gguad3<-sapply(seq(100000,200000,1000),guad3)
>plot(n,gguad3,xlab="Numero di simulazioni",
+ylab="Guadagno medio simulato",main="Paradosso di San Pietroburgo",
+p=0.6",type="l")
>abline(h=2*0.6/(2*0.6-1))
>
>guad4<-function(k){
+mean(2**(rgeom(k,prob=0.9)+1))}
>gguad4<-sapply(seq(100000,200000,1000),guad4)
>plot(n,gguad4,xlab="Numero di simulazioni",
+ylab="Guadagno medio simulato",main="Paradosso di San Pietroburgo",
+p=0.9",type="l")
>abline(h=2*0.9/(2*0.9-1))

```

La Figura 6.9 riporta il guadagno medio per alcune scelte di p . Ricordiamo che per $p = 0.4$ e per $p = 0.5$ il guadagno medio teorico è infinito, mentre per $p = 0.6$ è uguale a 6 e per $p = 0.9$ è uguale a 2.25 (linee orizzontali tracciate negli ultimi due grafici). Nel codice precedente è stata utilizzata la funzione `sapply(v, funz)` che permette di applicare la funzione indicata nel parametro `funz` a tutti gli elementi di una sequenza o di un vettore v restituendo un vettore di valori.

6.5 Distribuzione ipergeometrica

La distribuzione ipergeometrica interviene specificamente nella descrizione di estrazioni *senza reinserimento* oppure di estrazioni in blocco.

Si consideri l'esperimento che consiste nell'estrarre k biglie senza reinserimento da un'urna contenente $m + n$ biglie, di cui m sono bianche e n sono nere ($0 \leq k \leq m + n$) e si consideri l'evento

$$E_r = \{r \text{ delle } k \text{ biglie estratte sono bianche}\} \quad (r = 0, 1, \dots, k).$$

Facendo ricorso alla definizione classica di probabilità, si ha:

$$P(E_r) = \frac{\binom{m}{r} \binom{n}{k-r}}{\binom{m+n}{k}}, \quad (6.16)$$

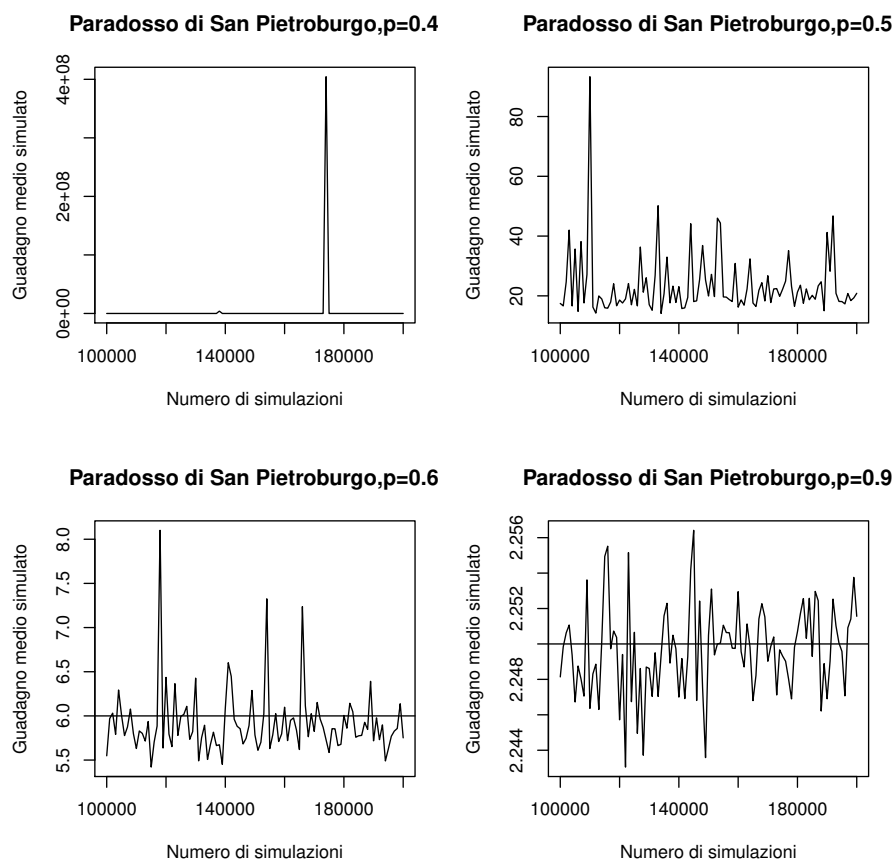


Figura 6.9: Guadagno medio simulato per diverse scelte di p . La linea orizzontale indica il guadagno medio teorico nel caso in cui $p > 0.5$.

dove il numeratore fornisce il numero di modi in cui si possono estrarre r delle m biglie bianche e $k-r$ delle n biglie nere presenti nell'urna, mentre il denominatore dà il numero di modi in cui si possono estrarre k delle $m+n$ biglie contenute nell'urna. Evidentemente deve essere $0 \leq r \leq m$ e $0 \leq k-r \leq n$, ossia

$$\max\{0, k-n\} \leq r \leq \min\{m, k\}.$$

Se si indica con X la variabile aleatoria che descrive il numero di biglie bianche estratte senza reinserimento, risulta $X = r$ se e solo se si verifica l'evento E_r ; pertanto risulta $P(X = r) = P(E_r)$ per $\max\{0, k-n\} \leq r \leq \min\{m, k\}$.

Definizione 6.4 Una variabile aleatoria X di funzione di probabilità

$$p_X(x) = \begin{cases} \frac{\binom{m}{x} \binom{n}{k-x}}{\binom{m+n}{k}}, & \max\{0, k-n\} \leq x \leq \min\{m, k\} \\ 0, & \text{altrimenti,} \end{cases} \quad (6.17)$$

con n, m e k interi tali che $0 \leq k \leq n+m$, è detta avere distribuzione ipergeometrica di parametri k, m, n .

Nel seguito con la notazione $X \sim \mathcal{I}(m, n, k)$ intenderemo che X è una variabile aleatoria avente distribuzione ipergeometrica di parametri m, n, k ; X sarà anche detta *variabile ipergeometrica*.

Il valore medio e la varianza della distribuzione ipergeometrica risultano essere:

$$E(X) = k \frac{m}{m+n}, \quad \text{Var}(X) = k \frac{m}{m+n} \frac{n}{m+n} \frac{m+n-k}{m+n-1}. \quad (6.18)$$

Se poniamo $p = m/(m+n)$, si nota che la media della distribuzione ipergeometrica coincide con la media di una variabile aleatoria binomiale $X \sim \mathcal{B}(k, p)$ e la varianza della distribuzione ipergeometrica è $(m+n-k)/(m+n-1)$ volte la varianza della distribuzione binomiale.

R permette di calcolare la funzione di probabilità, la funzione di distribuzione e i quantili di una variabile aleatoria ipergeometrica e anche di simulare tale variabile.

Si può richiedere ad R di eseguire direttamente il calcolo delle probabilità ipergeometriche utilizzando la funzione

`dhyper(x, m, n, k)`

Gli argomenti di tale funzione sono

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria ipergeometrica considerata;
- m indica il numero di palline bianche nell'urna;

- n indica il numero di palline nere nell'urna;
- k il numero di palline estratte dall'urna.

Ad esempio, se il numero di palline bianche nell'urna è $m = 12$, il numero di palline nere nell'urna è $n = 36$ e il numero di palline estratte dall'urna è $k = 5$, allora $\max(0, k - n) = 0$ e $\min(m, k) = 5$ e quindi le probabilità ipergeometriche possono essere così valutate:

```
> x<-0:5
> dhyper(x,12,36,5)
[1] 0.2201665125 0.4128122109 0.2752081406 0.0809435708
[5] 0.0104070305 0.0004625347
```

Si nota che la somma delle probabilità è unitaria.

Le seguenti linee di codice permettono di visualizzare le funzioni di probabilità ipergeometrica di Figura 6.10.

```
> par(mfrow=c(2,2))
> x<-0:5
> plot(x,dhyper(x,12,36,5),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="m=12,n=36,k=5")
>
> x<-0:5
> plot(x,dhyper(x,5,20,10),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="m=5,n=20,k=10")
>
> x<-0:20
> plot(x,dhyper(x,30,30,20),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="m=30,n=30,k=20")
>
> x<-0:20
> plot(x,dhyper(x,30,100,20),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="m=30,n=100,k=20")
```

Si nota che se $m = 12$, $n = 36$ e $k = 5$ risulta $\max(0, k - n) = 0$ e $\min(m, k) = 5$, da cui si ha $x = 0, 1, \dots, 5$. Se invece, $m = 5$, $n = 20$ e $k = 10$ risulta $\max(0, k - n) = 0$ e $\min(m, k) = 5$, da cui ancora si ha $x = 0, 1, \dots, 5$. Inoltre, se $m = 30$, $n = 30$ e $k = 20$ risulta $\max(0, k - n) = 0$ e $\min(m, k) = 20$, da cui ancora si ha $x = 0, 1, \dots, 20$. Infine, se $m = 30$, $n = 100$ e $k = 20$ risulta $\max(0, k - n) = 0$ e $\min(m, k) = 20$, da cui ancora si ha $x = 0, 1, \dots, 20$.

Si può richiedere ad R di eseguire direttamente il calcolo della funzione di distribuzione ipergeometrica utilizzando la funzione

```
phyper(x, m, n, k, lower.tail = TRUE)
```

Gli argomenti di tale funzione sono

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria ipergeometrica considerata;

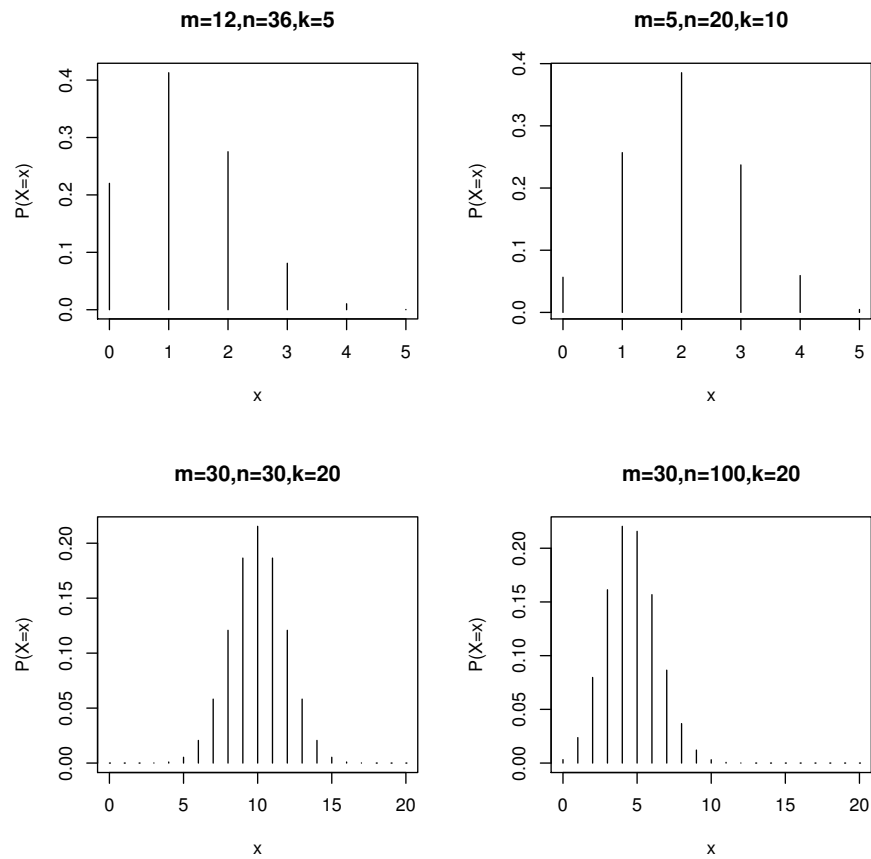


Figura 6.10: Rappresentazione della funzione di probabilità ipergeometrica per alcuni valori di m , n e k .

- m indica il numero di palline bianche nell'urna;
- n indica il numero di palline nere nell'urna;
- k il numero di palline estratte dall'urna;
- `lower.tail` se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

Ad esempio, se il numero di palline bianche nell'urna è $m = 12$, il numero di palline nere nell'urna è $n = 36$ e il numero di palline estratte dall'urna è $k = 5$ la distribuzione ipergeometrica può essere così valutata:

```
> x<-0:5
> phyper(x,12,36,5)
[1] 0.2201665 0.6329787 0.9081869 0.9891304 0.9995375 1.0000000
```

i cui risultati sono le probabilità:

$$P(X \leq x) = \sum_{n=0}^x P(X = n), \quad x = 0, 1, \dots, 5.$$

Inoltre, se $m = 12$, $n = 36$ e $k = 5$ le seguenti linee di codice

```
> x<-0:5
> phyper(x,12,36,5,lower.tail=FALSE)
[1] 0.7798334875 0.3670212766 0.0918131360 0.0108695652
[5] 0.0004625347 0.0000000000
```

mostrano le probabilità:

$$P(X > x) = 1 - P(X \leq x), \quad x = 0, 1, \dots, 5.$$

Le seguenti linee di codice permettono di visualizzare le funzioni di distribuzione ipergeometrica di Figura 6.11.

```
> par(mfrow=c(2,2))
> x<-0:5
> plot(x, phyper(x,12,36,5),
+ xlab="x", ylab=expression(P(X<=x)), ylim=c(0,1), type="s",
+ main="m=12, n=36, k=5")
>
> x<-0:5
> plot(x, phyper(x,5,20,10),
+ xlab="x", ylab=expression(P(X<=x)), ylim=c(0,1), type="s",
+ main="m=5, n=20, k=10")
>
> x<-0:20
> plot(x, phyper(x,30,30,20),
+ xlab="x", ylab=expression(P(X<=x)), ylim=c(0,1), type="s",
+ main="m=30, n=30, k=20")
>
> x<-0:20
> plot(x, phyper(x,30,100,20),
+ xlab="x", ylab=expression(P(X<=x)), ylim=c(0,1), type="s",
+ main="m=30, n=100, k=20")
```

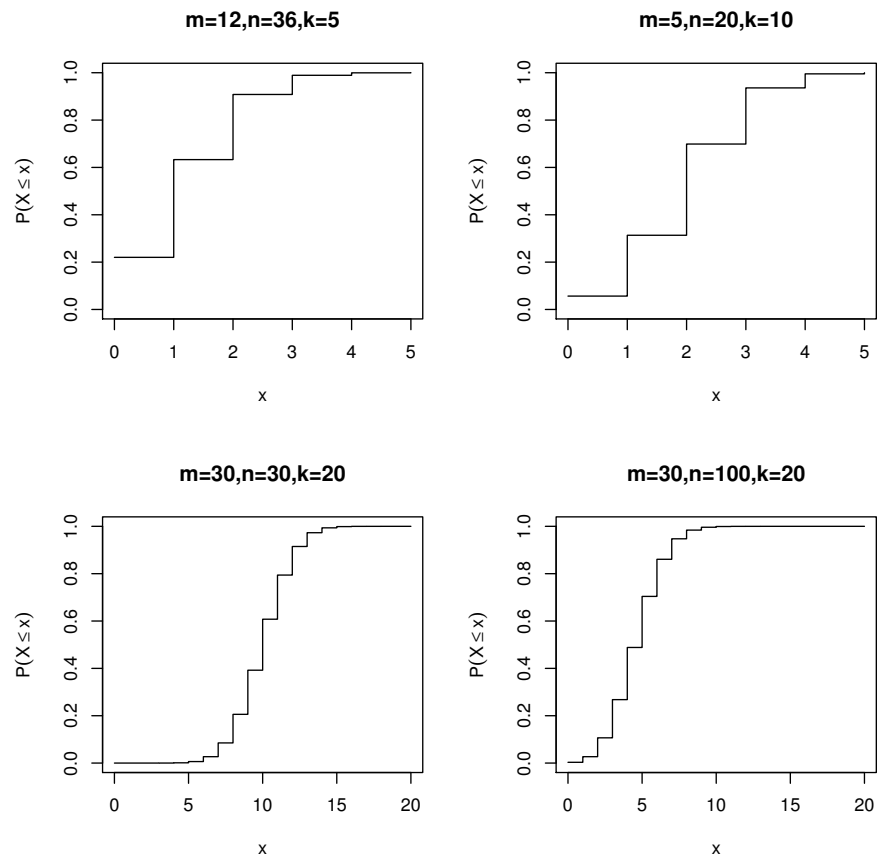


Figura 6.11: Rappresentazione della funzione di distribuzione ipergeometrica per alcuni valori di m , n e k .

È possibile calcolare il valore medio, la varianza, la deviazione standard e il coefficiente di variazione della distribuzione ipergeometrica attraverso la (6.18). Ad esempio, se $m = 5$, $n = 20$ e $k = 10$ si ha $E(X) = 2$, $\text{Var}(X) = 1$, $\sqrt{\text{Var}(X)} = 1$ e $\text{CV}(X) = 1/2$.

In R si possono calcolare anche i quantili (percentili) della distribuzione ipergeometrica attraverso la funzione

```
qhyper(z, m, n, k, lower.tail = TRUE)
```

dove

- z è il valore assunto (o i valori assunti) dalle probabilità relative al percentile $z \cdot 100$ -esimo;
- m indica il numero di palline bianche nell'urna;
- n indica il numero di palline nere nell'urna;
- k il numero di palline estratte dall'urna;
- `lower.tail` se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

Il risultato della funzione è il percentile $z \cdot 100$ -esimo, ossia il più piccolo numero intero r assunto dalla variabile aleatoria ipergeometrica X tale che

$$P(X \leq r) \geq z \quad \max\{0, k - n\} \leq r \leq \min\{m, k\}. \quad (6.19)$$

Ad esempio, se $m = 5$, $n = 20$ e $k = 10$ le seguenti linee di codice forniscono i quartili Q_0, Q_1, Q_2, Q_3, Q_4

```
> z<-c(0,0.25,0.5,0.75,1)
> qhyper(z,5,20,10)
[1] 0 1 2 3 5
```

che mostra che il primo quartile (25-esimo percentile) è $Q_1 = 1$, il secondo quartile o mediana (50-esimo percentile) è $Q_2 = 2$ e il terzo quartile (75-esimo percentile) è $Q_3 = 3$. Il minimo è $Q_0 = 0$ e il massimo è $Q_4 = 5$.

È possibile simulare in R la variabile aleatoria ipergeometrica generando una sequenza di numeri pseudocasuali mediante la funzione

```
rhyper(N, m, n, k)
```

dove

- N è lunghezza della sequenza da generare;
- m indica il numero di palline bianche nell'urna;
- n indica il numero di palline nere nell'urna;
- k il numero di palline estratte dall'urna;

Ad esempio, se desideriamo generare una sequenza di 25 numeri pseudocasuali simulando una variabile aleatoria ipergeometrica con $m = 5$, $n = 20$ e $k = 10$ si ha:

```
> sim<-rhyper(25,5,20,10)
> sim
[1] 3 2 3 3 3 2 2 0 2 2 0 3 2 2 2 1 1 2 2 3 1 1 2 2 3
> table(sim)
sim
 0   1   2   3
 2   4  12   7
> table(sim)/length(sim)
sim
 0       1       2       3
0.08 0.16 0.48 0.28
```

dove `table(sim)/length(sim)` fornisce le frequenze relative con cui i numeri $0, 1, \dots, 5$ si presentano nella sequenza generata. Occorre sottolineare che differenti esecuzioni conducono a sequenze pseudocasuali diverse.

Il codice seguente permette di confrontare la funzione di probabilità ipergeometrica teorica con quella simulata all'aumentare della lunghezza $N = 500, 5000, 50000$ della sequenza generata.

```
>par(mfrow=c(2,2))
>x<-0:5
>plot(x,dhyper(x,5,20,10),xlab="x",ylab="Probabilita'",type="h",
+main="m=5,n=20,k=10",xlim=c(0,5))
>
>sim1<-rhyper(500,5,20,10)
>plot(table(sim1)/length(sim1),xlab="x",type="h",
+ylab="Frequenza relativa",xlim=c(0,5),
+ylim=c(0,0.4),main="m=5,n=20,k=10,N=500")
>
>sim2<-rhyper(5000,5,20,10)
>plot(table(sim2)/length(sim2),xlab="x",type="h",
+ylab="Frequenza relativa",xlim=c(0,5),
+ylim=c(0,0.4),main="m=5,n=20,k=10,N=5000")
>
>sim3<-rhyper(50000,5,20,10)
>plot(table(sim3)/length(sim3),xlab="x",type="h",
+ylab="Frequenza relativa",xlim=c(0,5),
+ylim=c(0,0.4),main="m=5,n=20,k=10,N=50000")
```

Si nota che all'aumentare della lunghezza della sequenza generata il grafico delle frequenze relative si avvicina sempre di più al grafico della funzione di probabilità ipergeometrica.

Vogliamo ora mostrare che sotto opportune ipotesi la distribuzione ipergeometrica tende alla distribuzione binomiale.

Proposizione 6.2 *Sia X una variabile aleatoria ipergeometrica descrivente l'estrazione di k biglie senza reinserimento da un'urna contenente $m + n$ biglie, di cui m sono bianche e n sono nere ($0 \leq k \leq m + n$). Se m e $m + n$ divergono*

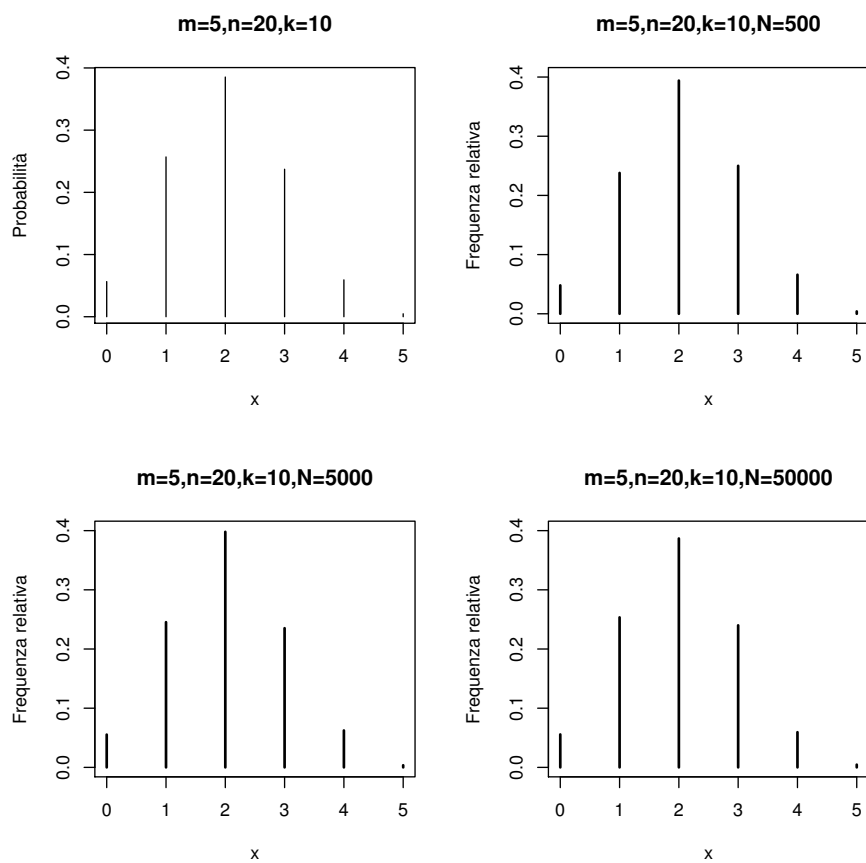


Figura 6.12: Confronto della funzione di probabilità ipergeometrica teorica e delle frequenze relative simulate per una variabile aleatoria ipergeometrica con $m = 5$, $n = 20$ e $k = 10$.

in maniera tale che $m/(m+n)$ converga ad un valore $p \in (0,1)$, allora

$$\lim_{\substack{m \rightarrow +\infty, m+n \rightarrow +\infty \\ m/(m+n) \rightarrow p}} p_X(x) = \binom{k}{x} p^x (1-p)^{k-x} \quad (x = 0, 1, \dots, k). \quad (6.20)$$

Con riferimento allo schema di estrazione che ha condotto alla formula (6.16), la Proposizione 6.2 comporta che se il numero m delle biglie bianche e il numero $m+n$ di biglie presenti nell'urna sono entrambi sufficientemente elevati in modo tale che il loro rapporto sia una costante p , allora la probabilità che x delle k biglie estratte senza reinserimento siano bianche è approssimabile con la medesima probabilità relativa al caso di estrazioni con reinserimento, essendo

$$\frac{\binom{m}{x} \binom{n}{k-x}}{\binom{m+n}{k}} \simeq \binom{k}{x} \left(\frac{m}{m+n} \right)^x \left(1 - \frac{m}{m+n} \right)^{k-x} \quad (x = 0, 1, \dots, k). \quad (6.21)$$

Il codice seguente permette di confrontare le probabilità presenti al primo membro ed al secondo membro della (6.21) per $k = 10$ e varie scelte di m, n tali che $p = m/(m+n) = 0.1$

```
>par(mfrow=c(2,2))
>x<-0:5
>plot(x,dhyper(x,2,18,10),
+ xlab="x",ylab="P(X=x)",type="h",ylim=c(0,0.6),
+ main="Ipergeometrica ,m=2,n=18,k=10")
>y1<-round(dhyper(x,2,18,10),3)
>text(x+0.04,dhyper(x,2,18,10)+0.03,y1)
>
>x<-0:5
>plot(x,dhyper(x,20,180,10),
+ xlab="x",ylab="P(X=x)",type="h",ylim=c(0,0.6),
+ main="Ipergeometrica ,m=20,n=180,k=10")
>y2<-round(dhyper(x,20,180,10),3)
>text(x+0.04,dhyper(x,20,180,10)+0.03,y2)
>
>x<-0:5
>plot(x,dhyper(x,200,1800,10),
+ xlab="x",ylab="P(X=x)",type="h",ylim=c(0,0.6),
+ main="Ipergeometrica ,m=200,n=1800,k=10")
>y3<-round(dhyper(x,200,1800,10),3)
>text(x+0.04,dhyper(x,200,1800,10)+0.03,y3)
>
>x<-0:5
>plot(x,dbinom(x,size=10,prob=0.1),
+ xlab="x",ylab="P(X=x)",type="h",ylim=c(0,0.6),
+ main="Binomiale ,k=10,p=0.1")
>y4<-round(dbinom(x,size=10,prob=0.1),3)
>text(x+0.04,dbinom(x,size=10,prob=0.1)+0.03,y4)
```

Come mostrato in Figura 6.13 l'approssimazione della distribuzione ipergeometrica con quella binomiale tende a migliorare al crescere di m e $m+n$ tali che $m/(m+n)$ è costante.

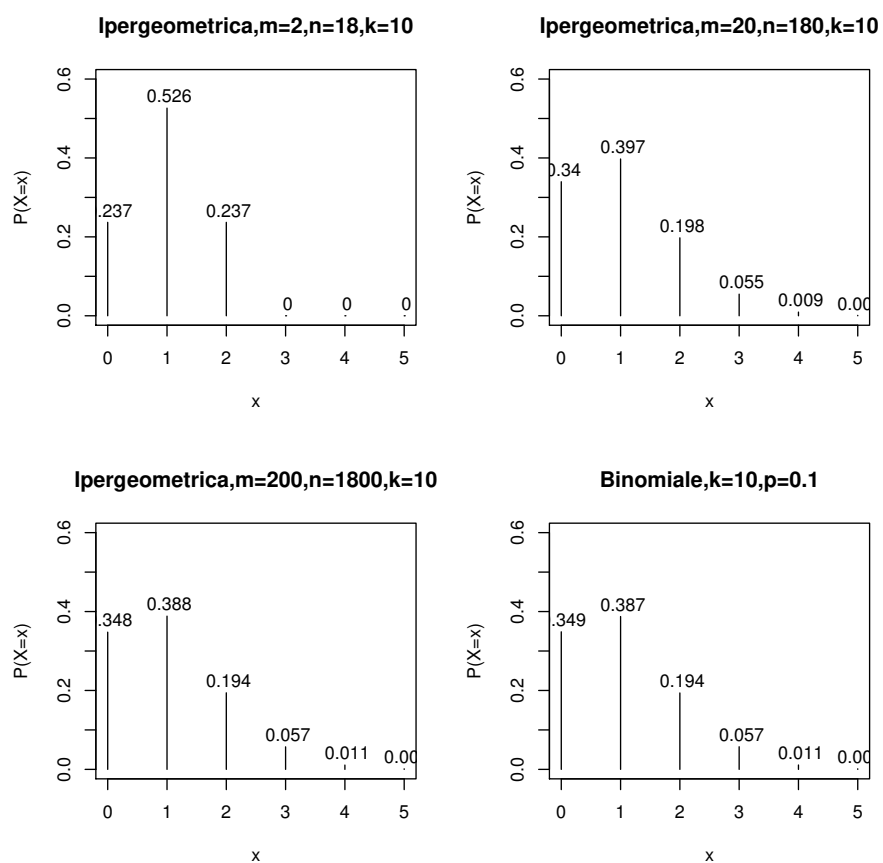


Figura 6.13: Confronto della funzione di probabilità ipergeometrica per $k = 10$ e varie scelte di m, n tali che $p = m/(n + m) = 0.1$ con la funzione di probabilità binomiale con $k = 10$ e $p = 0.1$.

Il codice seguente permette di visualizzare sullo stesso grafico le differenze tra la distribuzione ipergeometrica e la distribuzione binomiale.

```
>par(mfrow=c(2,2))
>x<-0:10
>matplot(x,data.frame(dbinom(x,size=10,prob=0.1),
+ dhyper(x,5,45,10)),pch=25,xlab="x",ylab="P(X=x)",
+ ylim=c(0,0.5),main="m=5,n=45,k=10")
>segments(x,dbinom(x,size=10,prob=0.1),x,dhyper(x,5,45,10))
>
>x<-0:10
>matplot(x,data.frame(dbinom(x,size=10,prob=0.1),
+ dhyper(x,10,90,10)),pch=25,xlab="x",ylab="P(X=x)",
+ ylim=c(0,0.5),main="m=10,n=90,k=10")
>segments(x,dbinom(x,size=10,prob=0.1),x,dhyper(x,10,90,10))
>
>x<-0:10
>matplot(x,data.frame(dbinom(x,size=10,prob=0.1),
+ dhyper(x,15,135,10)),pch=25,xlab="x",ylab="P(X=x)",
+ ylim=c(0,0.5),main="m=15,n=135,k=10")
>segments(x,dbinom(x,size=10,prob=0.1),x,dhyper(x,15,135,10))
>
>x<-0:10
>matplot(x,data.frame(dbinom(x,size=10,prob=0.1),
+ dhyper(x,30,270,10)),pch=25,xlab="x",ylab="P(X=x)",
+ ylim=c(0,0.5),main="m=30,n=270,k=10")
>segments(x,dbinom(x,size=10,prob=0.1),x,dhyper(x,30,270,10))
```

ed il relativo grafico è riportato in Figura 6.14.

Possiamo quindi concludere che quando m e $m+n$ sono sufficientemente grandi l'estrazione senza rimpiazzamento ottenuta utilizzando il modello ipergeometrico è praticamente equivalente all'estrazione con rimpiazzamento del modello binomiale.

6.6 Distribuzione di Poisson

La distribuzione di Poisson interviene spesso nella descrizione di alcuni fenomeni coinvolgenti qualche tipo di conteggio, quali il numero di chiamate telefoniche ricevute da un centralino in un fissato intervallo di tempo, il numero di particelle radioattive emesse per unità di tempo, il numero di microorganismi per unità di volume in un fluido, il numero di imperfezioni per unità di lunghezza di un cavo.

Definizione 6.5 Una variabile aleatoria X avente funzione di probabilità

$$p_X(x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda}, & x = 0, 1, \dots \quad (\lambda > 0), \\ 0, & \text{altrimenti} \end{cases} \quad (6.22)$$

è detta di distribuzione di Poisson di parametro λ .

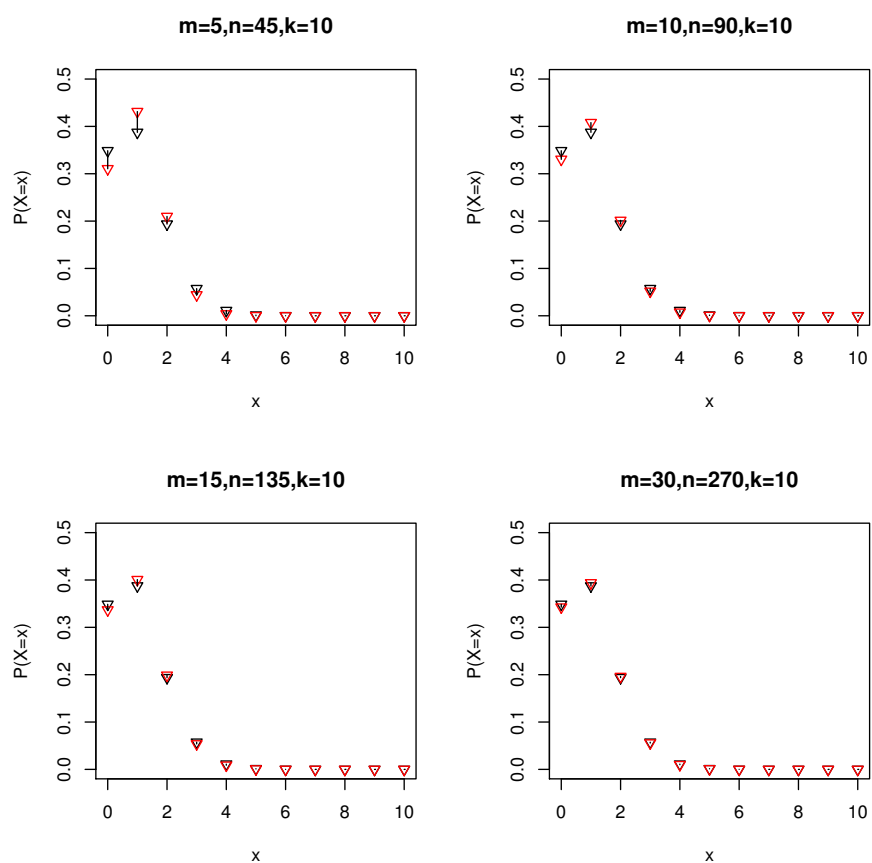


Figura 6.14: Differenze tra la funzione di probabilità ipergeometrica per $k = 10$ e varie scelte di m, n tali che $p = m/(n + m) = 0.1$ e la funzione di probabilità binomiale con $k = 10$ e $p = 0.1$.

Nel seguito con la notazione $X \sim \mathcal{P}(\lambda)$ si indicherà che X è una variabile aleatoria avente distribuzione di Poisson di parametro λ , o più semplicemente che X è una *variabile di Poisson* o *poissoniana*. Dalla (6.22) si ricava:

$$\frac{p_X(r)}{p_X(r-1)} = \frac{\lambda}{r} \quad (r = 1, 2, \dots), \quad (6.23)$$

così che le probabilità di Poisson (6.22) sono calcolabili in modo ricorsivo:

$$p_X(0) = e^{-\lambda}, \quad p_X(r) = \frac{\lambda}{r} p_X(r-1) \quad (r = 1, 2, \dots).$$

Per una variabile aleatoria di Poisson si ha:

$$E(X) = \lambda, \quad \text{Var}(X) = \lambda. \quad (6.24)$$

R permette di calcolare la funzione di probabilità, la funzione di distribuzione e i quantili di una variabile aleatoria di Poisson e anche di simulare tale variabile.

Si può richiedere ad R di eseguire direttamente il calcolo delle probabilità di Poisson utilizzando la funzione

```
dpois(x, lambda)
```

Gli argomenti di tale funzione sono

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria di Poisson considerata;
- λ vettore dei valori medi (non negativi).

Ad esempio, se $\lambda = 3$ le probabilità di Poisson per $x = 0, 1, \dots, 10$ possono essere così valutate:

```
> x<-0:10
> dpois(x,3)
[1] 0.0497870684 0.1493612051 0.2240418077 0.2240418077
[5] 0.1680313557 0.1008188134 0.0504094067 0.0216040315
[9] 0.0081015118 0.0027005039 0.0008101512
```

Il codice seguente permette di confrontare la funzione di probabilità di Poisson per alcune scelte di λ , il cui grafico è mostrato in Figura 6.15. Nel primo caso ($\lambda = 0.5$), $p_X(x)$ è strettamente decrescente, mentre per $\lambda = 2.5$ essa presenta un unico massimo in $x = 2$; nei rimanenti casi, essendo λ intero, $p_X(r)$ presenta due massimi le cui ascisse aumentano all'aumentare di λ , mentre le ordinate diminuiscono.

```
> par(mfrow=c(2,2))
> x<-0:5
> plot(x,dpois(x,lambda=0.5),
+ xlab="x",ylab="P(X=x)",type="h",main="lambda=0.5")
>
> x<-0:10
> plot(x,dpois(x,lambda=2.5),
```

```

+xlabel="x",ylab="P(X=x)",type="h",main="lambda=2.5")
>
>x<-0:10
>plot(x,dpois(x,lambda=3),
+xlabel="x",ylab="P(X=x)",type="h",main="lambda=3")
>
>x<-0:15
>plot(x,dpois(x,lambda=6),
+xlabel="x",ylab="P(X=x)",type="h",main="lambda=6")

```

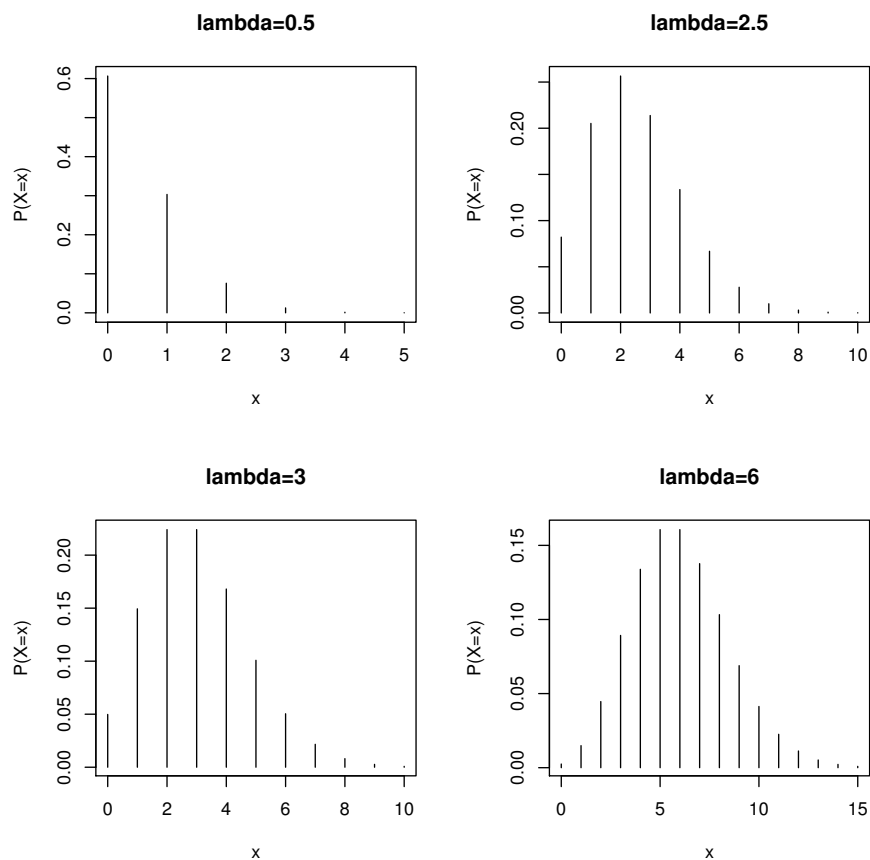


Figura 6.15: Rappresentazione della funzione di probabilità di Poisson per alcuni valori di λ .

Si può richiedere ad R di eseguire direttamente il calcolo della funzione di distribuzione di Poisson utilizzando la funzione

```
ppois(x, lambda, lower.tail = TRUE)
```

Gli argomenti di tale funzione sono

A.G. Nobile

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria di Poisson considerata;
- λ è il vettore dei valori medi (non negativi);
- `lower.tail` se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

Ad esempio, se $\lambda = 0.5$ la funzione di distribuzione di Poisson per $x = 0, 1, \dots, 8$ può essere così valutata:

```
> x<-0:8
> ppois(x,lambda=0.5)
[1] 0.6065307 0.9097960 0.9856123 0.9982484 0.9998279 0.9999858
[7] 0.9999990 0.9999999 1.0000000
```

mentre aggiungendo il parametro `lower.tail = FALSE` si può calcolare la $P(X > x)$. Le seguenti linee di codice permettono di visualizzare le funzioni di distribuzione di Figura 6.16.

```
>par(mfrow=c(2,2))
>x<-0:5
>plot(x,ppois(x,lambda=0.5),
+xlabs="x",ylab=expression(P(X<=x)),ylim=c(0,1),type="s",
+main="lambda=0.5")
>
>x<-0:10
>plot(x,ppois(x,lambda=2.5),
+xlabs="x",ylab=expression(P(X<=x)),ylim=c(0,1),type="s",
+main="lambda=2.5")
>
>x<-0:10
>plot(x,ppois(x,lambda=3),
+xlabs="x",ylab=expression(P(X<=x)),ylim=c(0,1),type="s",
+main="lambda=3")
>
>x<-0:15
>plot(x,ppois(x,lambda=6),
+xlabs="x",ylab=expression(P(X<=x)),ylim=c(0,1),type="s",
+main="lambda=6")
```

È possibile calcolare il valore medio, la varianza, la deviazione standard e il coefficiente di variazione della distribuzione ipergeometrica attraverso la (6.24). Ad esempio, se $\lambda = 3$ si ha $E(X) = 3$, $\text{Var}(X) = 3$, $\sqrt{\text{Var}(X)} = \sqrt{3}$ e $\text{CV}(X) = 1/\sqrt{3}$.

In R si possono calcolare anche i quantili (percentili) della distribuzione di Poisson attraverso la funzione

```
qpois(z, lambda, lower.tail = TRUE)
```

dove

- z è il valore assunto (o i valori assunti) dalle probabilità relative al percentile $z \cdot 100$ -esimo;

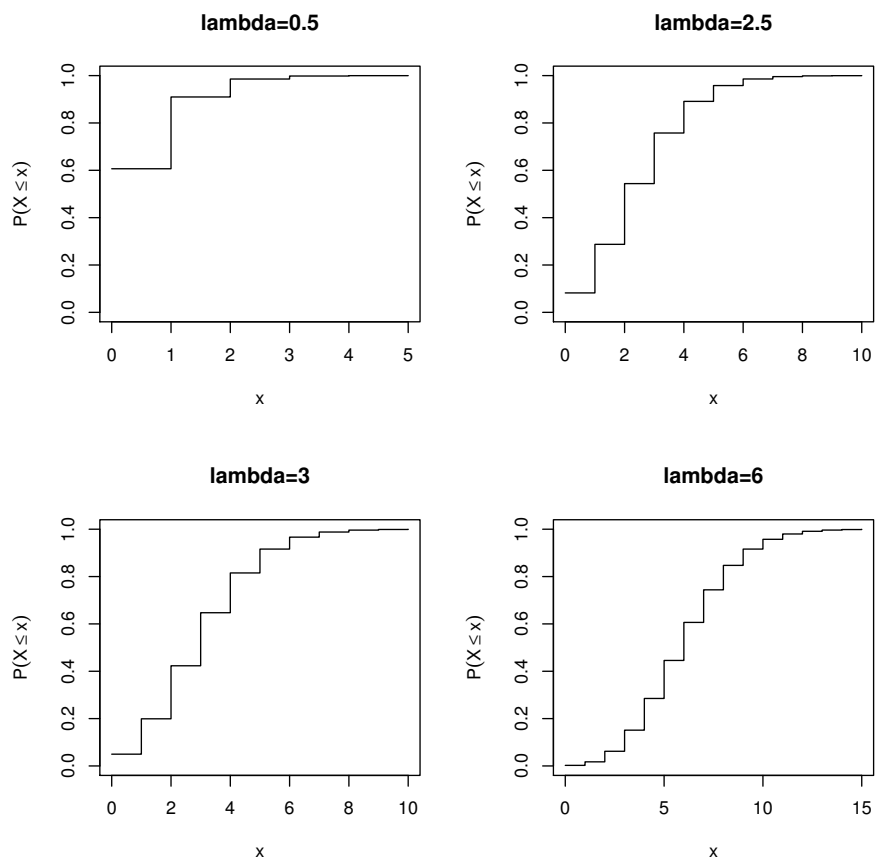


Figura 6.16: Rappresentazione della funzione di distribuzione di Poisson per alcuni valori di λ .

- `lambda` è il vettore dei valori medi (non negativi);
- `lower.tail` se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

Il risultato della funzione è il percentile $z \cdot 100$ -esimo, ossia il più piccolo numero intero k assunto dalla variabile aleatoria di Poisson X tale che

$$P(X \leq k) \geq z \quad (k = 0, 1, \dots). \quad (6.25)$$

Ad esempio, se $\lambda = 3$ le seguenti linee di codice forniscono i quartili Q_0, Q_1, Q_2, Q_3, Q_4

```
>z<-c(0,0.25,0.5,0.75,1)
> qpois(z,lambda=3)
[1] 0 2 3 4 Inf
```

che mostra che il primo quartile (25-esimo percentile) è $Q_1 = 2$, il secondo quartile o mediana (50-esimo percentile) è $Q_2 = 3$ e il terzo quartile (75-esimo percentile) è $Q_3 = 4$. Il minimo è $Q_0 = 0$ e il massimo è $Q_4 = \infty$.

È possibile simulare in R la variabile aleatoria di Poisson generando una sequenza di numeri pseudocasuali mediante la funzione

```
rpois(N,lambda)
```

dove

- `N` è lunghezza della sequenza da generare;
- `lambda` è il vettore dei valori medi (non negativi);

Ad esempio, se desideriamo generare una sequenza di 50 numeri pseudocasuali simulando una variabile aleatoria di Poisson di valor medio $\lambda = 3$ si ha:

```
> sim<-rpois(50,lambda=3)
> sim
[1] 1 1 4 0 3 0 2 1 5 3 2 1 5 2 2 2 5 1 4 2 1 4 2 3 5 4 3 4 0 3 3 1
[33] 7 2 1 2 2 2 3 4 2 1 2 5 1 2 1 4 5 3
> table(sim)
sim
 0  1  2  3  4  5  7
 3 11 14  8  7  6  1
> table(sim)/length(sim)
sim
 0    1    2    3    4    5    7
0.06 0.22 0.28 0.16 0.14 0.12 0.02
```

dove `table(sim)/length(sim)` fornisce le frequenze relative con cui i numeri $0, 1, \dots$, si presentano nella sequenza generata. Occorre sottolineare che differenti esecuzioni conducono a sequenze pseudocasuali diverse.

Il codice seguente permette di confrontare la funzione di probabilità di Poisson teorica con quella simulata all'aumentare della lunghezza $N = 500, 5000, 50000$ della sequenza generata.

```

>par(mfrow=c(2,2))
>x<-0:10
>plot(x,dpois(x,lambda=3),xlab="x",ylab="Probabilita'",type="h",
+main="lambda=3",xlim=c(0,10),ylim=c(0,0.25))
>
>sim1<-rpois(500,lambda=3)
>plot(table(sim1)/length(sim1),xlab="x",type="h",
+ylab="Frequenza relativa",xlim=c(0,10),ylim=c(0,0.25),
+main="lambda=3,N=500")
>
>sim2<-rpois(5000,lambda=3)
>plot(table(sim2)/length(sim2),xlab="x",type="h",
+ylab="Frequenza relativa",xlim=c(0,10),ylim=c(0,0.25),
+main="lambda=3,N=5000")
>
>sim3<-rpois(50000,lambda=3)
>plot(table(sim3)/length(sim3),xlab="x",type="h",
+ylab="Frequenza relativa",xlim=c(0,10),ylim=c(0,0.25),
+main="lambda=3,N=50000")

```

Si nota che all'aumentare della lunghezza della sequenza generata il grafico delle frequenze relative si avvicina sempre di più al grafico della funzione di probabilità di Poisson.

6.6.1 Approssimazione della distribuzione binomiale con la distribuzione di Poisson

La distribuzione di Poisson è spesso detta *degli eventi rari* o *dei piccoli numeri* poiché, come si vedrà qui di seguito, essa si rivela utile per trattare i cosiddetti eventi rari di schemi binomiali in cui la probabilità di successo in ogni singola prova è molto piccola mentre il numero di prove è molto grande, come spesso accade in numerosi fenomeni biologici (colonie di batteri, mutazioni genetiche), assicurativi (incidenti aerei, incendi), industriali (controllo statistico della qualità di prodotti), ecc.

Proposizione 6.3 *Sia X_1, X_2, \dots una successione di variabili aleatorie con $X_n \sim \mathcal{B}(n, p_n)$. Se al divergere di n , p_n tende a zero in modo tale che $np_n \rightarrow \lambda$, con $\lambda > 0$, allora*

$$\lim_{\substack{n \rightarrow +\infty, \\ np_n \rightarrow \lambda}} p_{X_n}(k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad (k = 0, 1, \dots). \quad (6.26)$$

Un'immediata conseguenza della Proposizione 6.3 è che se il numero n di prove ripetute di Bernoulli è elevato e se la probabilità di successo p in ogni prova è piccola, allora è possibile utilizzare la seguente approssimazione:

$$\binom{n}{k} p^k (1-p)^{n-k} \simeq \frac{(np)^k}{k!} e^{-np} \quad (k = 0, 1, \dots). \quad (6.27)$$

È opportuno menzionare che la distribuzione di Poisson costituisce una buona approssimazione della distribuzione binomiale quando nella (6.27) si ha $n \geq 20$ e $p \leq 0.05$; per $n \geq 100$ e $np \leq 10$ l'approssimazione diviene poi eccellente.

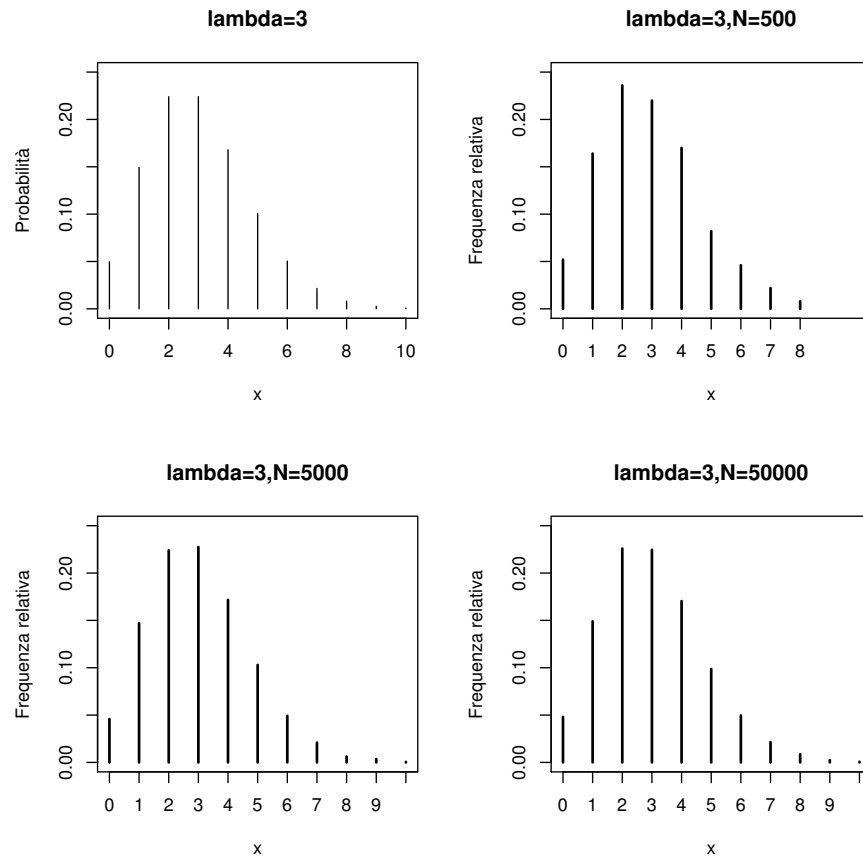


Figura 6.17: Confronto della funzione di probabilità di Poisson teorica e delle frequenze relative simulate per una variabile aleatoria di Poisson $X \sim \mathcal{P}(3)$.

Il codice seguente permette di confrontare le probabilità presenti al primo membro ed al secondo membro della (6.27) per $\lambda = 2$ e varie scelte di n e p tali che $np = 2$.

```
>par(mfrow=c(2,2))
>x<-0:6
>plot(x,dbinom(x,size=10,prob=0.2),
+xlabs="x",ylab="P(X=x)",type="h",ylim=c(0,0.35),
+main="Binomiale,n=10,p=0.2")
>y1<-round(dbinom(x,size=10,prob=0.2),3)
>text(x+0.04,dbinom(x,size=10,prob=0.2)+0.03,y1)
>
>x<-0:6
>plot(x,dbinom(x,size=50,prob=0.04),
+xlabs="x",ylab="P(X=x)",type="h",ylim=c(0,0.35),
+main="Binomiale,n=50,p=0.04")
>y2<-round(dbinom(x,size=50,prob=0.04),3)
>text(x+0.04,dbinom(x,size=50,prob=0.04)+0.03,y2)
>
>x<-0:6
>plot(x,dbinom(x,size=100,prob=0.02),
+xlabs="x",ylab="P(X=x)",type="h",ylim=c(0,0.35),
+main="Binomiale,n=100,p=0.02")
>y3<-round(dbinom(x,size=100,prob=0.02),3)
>text(x+0.04,dbinom(x,size=100,prob=0.02)+0.03,y3)
>
>x<-0:6
>plot(x,dpois(x,lambda=2),
+xlabs="x",ylab="P(X=x)",type="h",ylim=c(0,0.35),
+main="Poisson,lambda=2")
>y4<-round(dpois(x,lambda=2),3)
>text(x+0.04,dpois(x,lambda=2)+0.03,y4)
```

Come mostrato in Figura 6.18 l'approssimazione della distribuzione binomiale con quella di Poisson tende a migliorare al crescere di n e al diminuire di p in maniera tale che $np = \lambda$ sia costante. Il codice seguente permette di visualizzare sullo stesso grafico le differenze tra la distribuzione binomiale e la distribuzione di Poisson.

```
>par(mfrow=c(2,2))
>x<-0:10
>matplot(x,data.frame(dbinom(x,size=10,prob=0.2),
+dpois(x,lambda=2)),pch=25,xlab="x",ylab="P(X=x)",
+ylim=c(0,0.3),main="n=10,p=0.2")
>segments(x,dbinom(x,size=10,prob=0.2),x,dpois(x,lambda=2))
>
>x<-0:10
>matplot(x,data.frame(dbinom(x,size=25,prob=0.08),
+dpois(x,lambda=2)),pch=25,xlab="x",ylab="P(X=x)",
+ylim=c(0,0.3),main="n=25,p=0.08")
>segments(x,dbinom(x,size=25,prob=0.08),x,dpois(x,lambda=2))
>
>x<-0:10
>matplot(x,data.frame(dbinom(x,size=50,prob=0.04),
+dpois(x,lambda=2)),pch=25,xlab="x",ylab="P(X=x)",
+ylim=c(0,0.3),main="n=50,p=0.04")
```

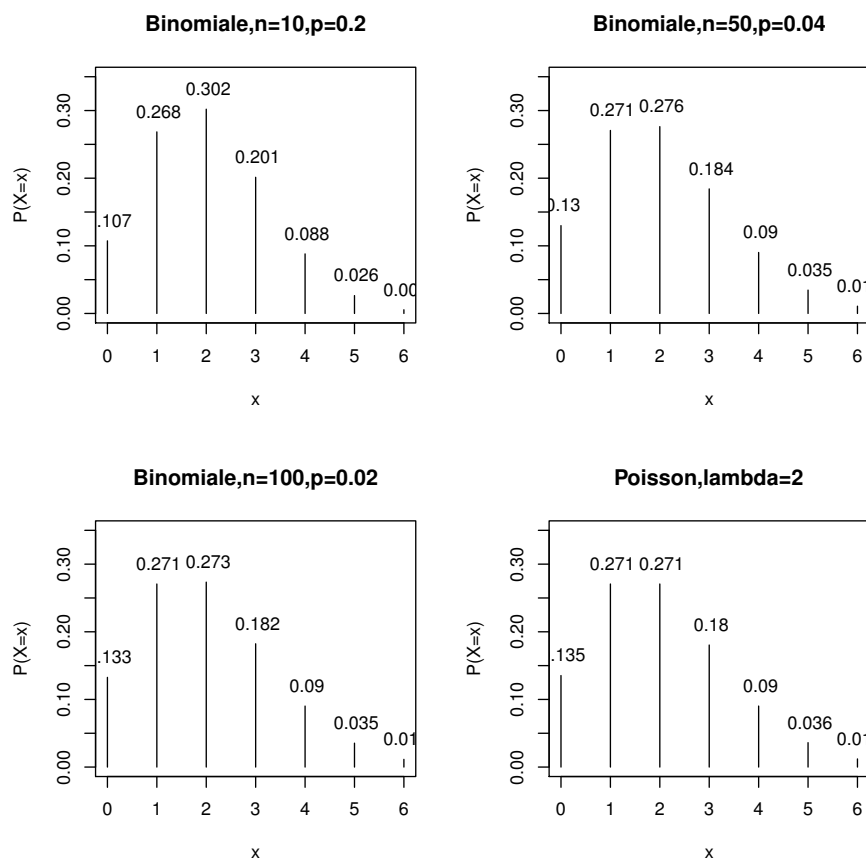


Figura 6.18: Confronto della funzione di probabilità binomiale per varie scelte di n e p tali che $np = 2$ con la funzione di probabilità di Poisson con $\lambda = 2$.

```

>segments(x,dbinom(x,size=50,prob=0.04),x,dpois(x,lambda=2))
>
>x<-0:10
>matplot(x,data.frame(dbinom(x,size=100,prob=0.02),
+dpois(x,lambda=2)),pch=25,xlab="x",ylab="P(X=x)",
+ylim=c(0,0.3),main="n=100,p=0.02")
>segments(x,dbinom(x,size=100,prob=0.02),x,dpois(x,lambda=2))

```

ed il relativo grafico è riportato in Figura 6.19.

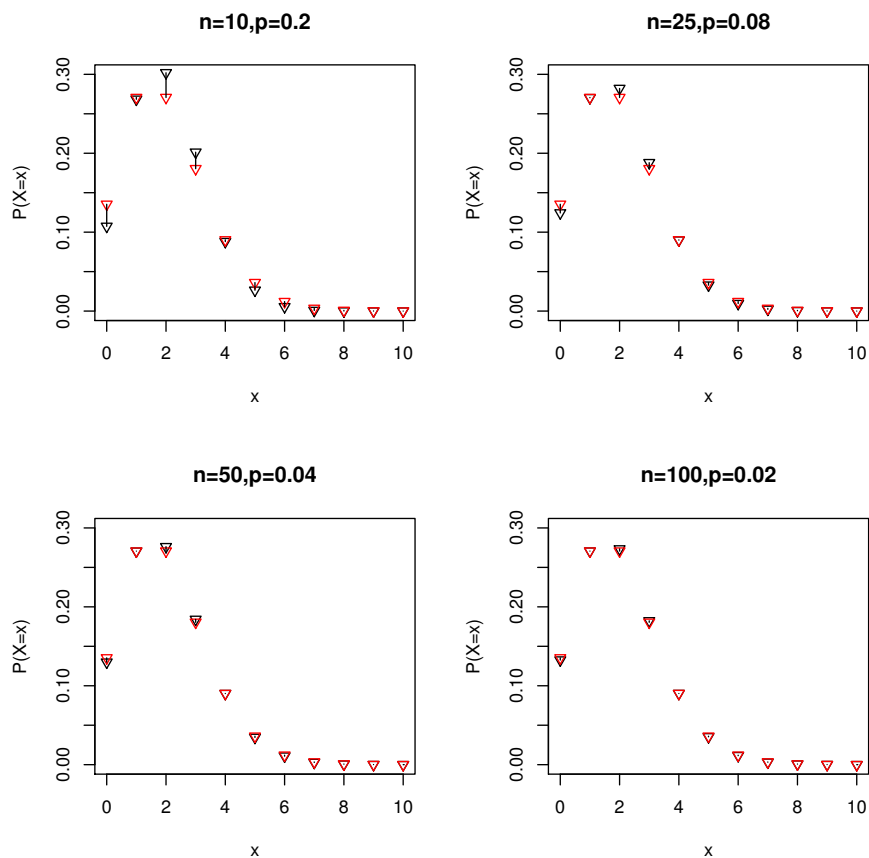


Figura 6.19: Differenze tra la funzione di probabilità binomiale per n e p tali che $np = 2$ e la funzione di probabilità di Poisson con $\lambda = 2$.

È evidente dalla Figura 6.19 che l'approssimazione della distribuzione binomiale con quella di Poisson migliora al crescere di n e al decrescere di p .

Esempio 6.3 Si supponga che la probabilità che un autoveicolo si guasti all'interno di un certo tunnel sia 0.0004. Si desidera calcolare la probabilità che di 1000 autoveicoli che attraversano il tunnel se ne guasti al più uno.

A.G. Nobile

Sia X una variabile aleatoria descrivente il numero di autoveicoli che si guastano all'interno del tunnel. Evidentemente $X \sim \mathcal{B}(1000, 0.0004)$, ossia X ha distribuzione binomiale di parametri $n = 1000$ e $p = 0.0004$. Pertanto, considerato l'evento $A = \{\text{si guasta al più un autoveicolo ogni mille che attraversano il tunnel}\}$, si ha:

$$P(A) = p_X(0) + p_X(1) = \binom{1000}{0} (0.0004)^0 (1 - 0.0004)^{1000} + \binom{1000}{1} (0.0004)^1 (1 - 0.0004)^{999} = 0.9385$$

Nel caso in esame, essendo $n = 1000$ e $np = 1000 \cdot 0.0004 = 0.4$, l'approssimazione di Poisson della distribuzione binomiale è ottima; infatti, ponendo $\lambda = 0.4$ si ha:

$$P(A) \simeq \frac{\lambda^0}{0!} e^{-\lambda} + \frac{\lambda^1}{1!} e^{-\lambda} = 1.4 \cdot e^{-0.4} = 0.9384$$

che differisce dal valore precedente sulla quarta cifra decimale. Ciò è confermato usando R:

```
>dbinom(0,1000,0.0004)+dbinom(1,1000,0.0004)
[1] 0.9384803
> dpois(0,0.4)+dpois(1,0.4)
[1] 0.938448
```

Esempio 6.4 Supponiamo che il numero di errori di battitura al minuto sia descritto da una variabile aleatoria di Poisson di valore medio 0.0001 (numero medio di errori di battitura al minuto). Considerando un intervallo di k minuti, possiamo utilizzare una distribuzione di Poisson con $\lambda = 0.0001 \cdot k$. Ciò è dovuto alla circostanza che la somma di variabili aleatorie di Poisson indipendenti è caratterizzata ancora da una distribuzione di Poisson il cui valore medio è la somma dei valori medi delle singole variabili di Poisson.

Desideriamo scegliere il più piccolo valore di k tale che la probabilità di non commettere errori in un intervallo di k minuti sia inferiore a 0.0005, ossia

$$P(\text{non commettere errori in } k \text{ minuti}) = e^{-0.0001 \cdot k} < 0.0005.$$

Ciò è equivalente a richiedere che

$$P(\text{si verifica almeno un errore in } k \text{ minuti}) = 1 - e^{-0.0001 \cdot k} > 0.9995.$$

è molto alta. Occorre quindi scegliere il più piccolo valore di k tale che

$$k \geq -\frac{\ln 0.0005}{0.0001} = 76009.02.$$

Il codice seguente

A.G. Nobile

```
k<-seq(50000,100000,1000)
> y<-exp(-0.0001*k)
> plot(k,y,xlab="k",ylab="Probabilita' di non commettere errori")
> abline(h=0.0005)
> ceiling(-log(0.0005)/0.0001)
[1] 76010
> 76010/60
[1] 1266.833
```

permette di ottenere il grafico in Figura 6.20. Ricordiamo che `ceiling(x)` fornisce il più piccolo intero maggiore di x .

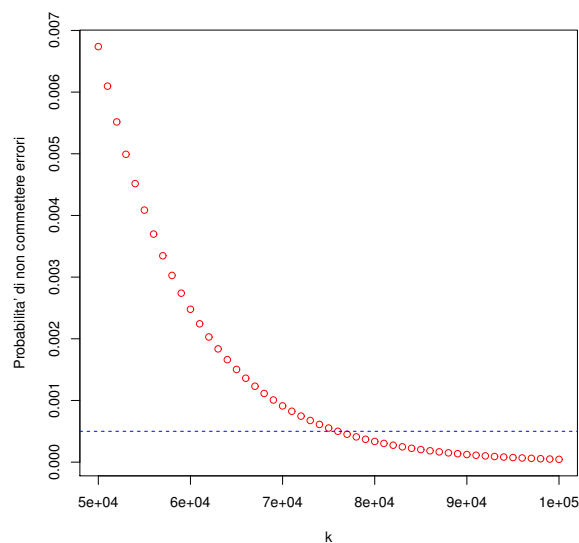


Figura 6.20: Probabilità di non commettere errori in k minuti.

Osservando il grafico si nota che la probabilità di non commettere errori in k minuti è una funzione decrescente in k e quindi occorreranno 76010 minuti, ossia circa 1267 ore, per ottenere una probabilità di non commettere errori inferiore a 0.0005.

Capitolo 7

Variabili aleatorie continue con R

7.1 Introduzione

Il sistema R mette a disposizione per ciascuna delle principali distribuzioni di probabilità continue:

- la funzione densità di probabilità;
- la funzione di distribuzione;
- la funzioni quantili;
- la funzione che simula tale variabile aleatoria mediante la generazione di numeri pseudocasuali;

Tutte queste funzioni utilizzano nomi che iniziano con una particolare lettera dell'alfabeto, in modo da indicare il tipo di funzione a cui fa riferimento, seguita dal nome della distribuzione teorica scelta. La particolare lettera dell'alfabeto può essere:

d calcola la densità di probabilità di una variabile aleatoria in uno specifico punto o in un insieme di punti;

p calcola la funzione di distribuzione di una variabile aleatoria in uno specifico punto o in un insieme di punti;

q calcola le funzioni quantili;

r calcola la funzione che simula una variabile aleatoria mediante la generazione di numeri pseudocasuali.

In questo capitolo considereremo le seguenti distribuzioni continue:

- distribuzione uniforme;
- distribuzione esponenziale;
- distribuzione normale;
- distribuzione chi-quadrato;
- distribuzione di Student.

7.2 Distribuzione uniforme

Definizione 7.1 Siano a e b numeri reali tali che $a < b$. Una variabile aleatoria X di funzione di distribuzione

$$F_X(x) = P(X \leq x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x < b \\ 1, & x \geq b \end{cases} \quad (7.1)$$

e corrispondente densità di probabilità

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{altrimenti} \end{cases} \quad (7.2)$$

si dice uniformemente distribuita o equidistribuita nell'intervallo (a, b) .

Si noti che la (7.2) coincide con $dF_X(x)/dx$ laddove $F_X(x)$ è derivabile ed è posta arbitrariamente uguale a zero per $x = a$ e $x = b$.

La densità di probabilità uniforme traduce nel continuo il concetto di equiprobabilità che è alla base della distribuzione uniforme discreta. Infatti, anche se la probabilità $P(X = x)$ è nulla per ciascun punto x , la densità di probabilità uniforme assegna valori uguali a tutti gli intervalli di uguale ampiezza che vengono scelti in (a, b) .

Nel seguito, con la notazione $X \sim \mathcal{U}(a, b)$ intenderemo che X è una variabile aleatoria avente distribuzione uniforme nell'intervallo (a, b) ; X sarà anche detta *variabile uniforme*. Per una variabile aleatoria uniforme $X \sim \mathcal{U}(a, b)$ si ha:

$$E(X) = \frac{a+b}{2}, \quad E(X^2) = \frac{a^2 + ab + b^2}{3}, \quad \text{Var}(X) = \frac{(b-a)^2}{12}.$$

R permette di calcolare la densità di probabilità, la funzione di distribuzione e i quantili di una variabile aleatoria uniforme e anche di simulare tale variabile.

Si può richiedere ad R di eseguire direttamente il calcolo della densità uniforme utilizzando la funzione

```
dunif(x, min=a, max=b)
```


Gli argomenti di tale funzione sono

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria uniforme;
- \min e \max sono il minimo e il massimo dell'intervallo in cui la densità uniforme è positiva; se il minimo ed il massimo non sono specificati essi per default assumono i valori 0 e 1.

Per calcolare la funzione di distribuzione invece utilizziamo la funzione

```
punif(x, min=a, max=b, lower.tail = TRUE)
```

Gli argomenti di tale funzione sono

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria uniforme;
- \min e \max sono il minimo e il massimo dell'intervallo in cui la densità uniforme è positiva;
- lower.tail se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

Ad esempio per rappresentare la densità di probabilità e la funzione di distribuzione di una variabile aleatoria $X \sim \mathcal{U}(0, 1)$ utilizziamo il codice

```
>par(mfrow=c(1,2))
>curve(dunif(x,min=0,max=1),from=-1, to=2,xlab="x",
+ylab="f(x)",main="a=0,b=1")
>
>curve(punif(x,min=0,max=1),from=-1, to=2,xlab="x",
+ylab=expression(P(X<=x)),main="a=0,b=1")
```

che produce il grafico di Figura 7.1.

Invece per rappresentare la densità di probabilità e la funzione di distribuzione di una variabile aleatoria $X \sim \mathcal{U}(4, 9)$ utilizziamo il codice

```
>par(mfrow=c(1,2))
>curve(dunif(x,min=4,max=9),from=3, to=10,xlab="x",
+ylab="f(x)",main="a=4,b=9")
>
>curve(punif(x,min=4,max=9),from=3, to=10,xlab="x",
+ylab=expression(P(X<=x)),main="a=4,b=9")
```

ottenendo il grafico di Figura 7.2.

La probabilità che la variabile aleatoria $X \sim \mathcal{U}(4, 9)$ assuma valori nell'intervallo (6,8) è

$$P(6 < X < 8) = (8 - 6) \cdot \frac{1}{9 - 4} = \frac{2}{5} = 0.4,$$

e corrisponde all'area del rettangolo visualizzato in Figura 7.3 visualizzata tramite il seguente codice:

A.G. Nobile

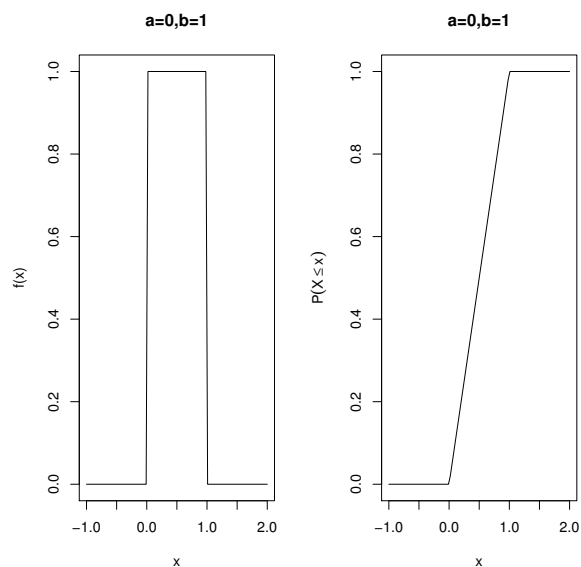


Figura 7.1: Rappresentazione della densità e della funzione di distribuzione uniforme nell'intervallo $(0, 1)$.

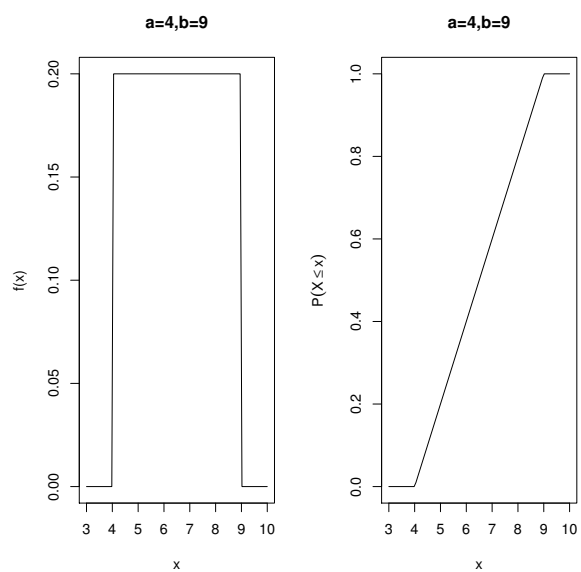


Figura 7.2: Rappresentazione della densità e della funzione di distribuzione uniforme nell'intervallo $(4, 9)$.

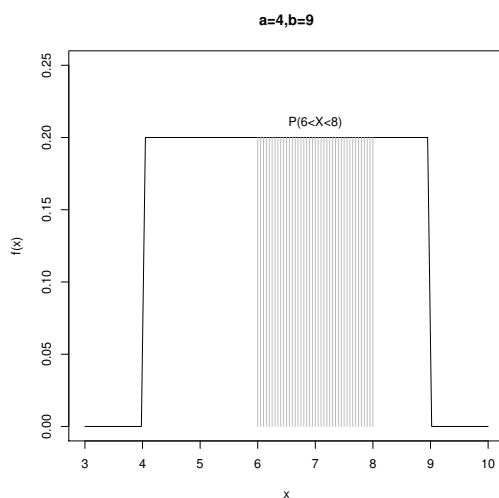


Figura 7.3: Rappresentazione della probabilità $P(6 < X < 8)$ per una variabile uniforme nell'intervallo $(4, 9)$.

```
>curve(dunif(x,min=4,max=9),from=3, to=10,xlab="x",
+ylab="f(x)",main="a=4,b=9",ylim=c(0,0.25))
>x<-seq(6,8,0.05)
>lines(x,dunif(x,min=4,max=9),type="h",col="grey")
>text(7.0,0.21,"P(6<X<8)")
>
> punif(8,min=4,max=9)-punif(6,min=4,max=9)
[1] 0.4
```

In R si possono calcolare anche i quantili (percentili) della distribuzione uniforme nell'intervallo (a, b) attraverso la funzione

```
qunif(z, min=a, max=b, lower.tail = TRUE)
```

dove

- z è il valore assunto (o i valori assunti) dalle probabilità relative al percentile $z \cdot 100$ -esimo;
- \min e \max sono il minimo e il massimo dell'intervallo in cui la densità uniforme è positiva;
- `lower.tail` se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

Il risultato della funzione è il percentile $z \cdot 100$ -esimo, ossia il più piccolo numero x assunto dalla variabile aleatoria uniforme X tale che

$$P(X \leq x) \geq z \quad (a < x < b). \quad (7.3)$$

Ad esempio, se si considera la variabile $X \sim \mathcal{U}(4, 9)$ le seguenti linee di codice forniscono i quartili Q_0, Q_1, Q_2, Q_3, Q_4

```
>z<-c(0,0.25,0.5,0.75,1)
>qunif(z, min=4, max=9)
[1] 4.00 5.25 6.50 7.75 9.00
```

che mostra che il primo quartile (25-esimo percentile) è $Q_1 = 5.25$, il secondo quartile o mediana (50-esimo percentile) è $Q_2 = 6.50$ e il terzo quartile (75-esimo percentile) è $Q_3 = 7.75$. Il minimo è $Q_0 = 4$ e il massimo è $Q_4 = 9$. Infatti, ricordando che $P(X \leq x) = (x - a)/(b - a) = (x - 4)/5$ se $4 \leq x < 9$ si ha

$$\begin{aligned} P(X \leq x) = \frac{x-4}{5} \geq 0.25 &\iff x \geq 4 + 5 \cdot 0.25 = 5.25, \\ P(X \leq x) = \frac{x-4}{5} \geq 0.50 &\iff x \geq 4 + 5 \cdot 0.50 = 6.50, \\ P(X \leq x) = \frac{x-4}{5} \geq 0.75 &\iff x \geq 4 + 5 \cdot 0.75 = 7.75. \end{aligned}$$

È possibile simulare in R la variabile aleatoria uniforme nell'intervallo (a, b) generando una sequenza di numeri pseudocasuali mediante la funzione

```
runif(N, min=a, max=b)
```

dove

- N è lunghezza della sequenza da generare;
- \min e \max sono il minimo e il massimo dell'intervallo in cui la densità uniforme è positiva.

Ad esempio, se desideriamo generare una sequenza di 10000 numeri pseudocasuali simulando una variabile aleatoria uniforme $X \sim \mathcal{U}(4, 9)$ si ha:

```
> sim<-runif(10000, min=4, max=9)
> mean(sim)
[1] 6.482359
> var(sim)
[1] 2.062938
```

che si avvicinano al valore medio teorico $E(X) = (a + b)/2 = 6.5$ e alla varianza teorica $\text{Var}(X) = (b - a)^2/12 = 25/12 = 2.083$.

Esempio 7.1 Due clienti A e B entrano contemporaneamente in un supermercato. Supponendo che il tempo impiegato per fare la spesa sia una variabile aleatoria uniformemente distribuita nell'intervallo $(10, 20)$ per A e nell'intervallo $(15, 25)$ per B e che siano tra loro indipendenti, si desidera determinare la probabilità che B finisca la spesa dopo di A.

Denotiamo con $X \sim \mathcal{U}(10, 20)$ la variabile aleatoria che descrive il tempo per fare la spesa di A e con $Y \sim \mathcal{U}(15, 25)$ la variabile aleatoria che descrive il tempo per fare la spesa di B. La probabilità richiesta può essere così calcolata

$$P(Y > X) = \int \int_{\mathcal{D}} f_{XY}(x, y) \, dx \, dy = \int \int_{\mathcal{D}} f_X(x) f_Y(y) \, dx \, dy$$

dove il dominio è così definito:

$$\mathcal{D} = \{(x, y) : 10 < x < 20, 15 < y < 25, y > x\}.$$

Il seguente codice permette di visualizzare il dominio (vedi Figura 7.4):

```
>plot(10:25,10:25,xlab="x",ylab="y",type="l")
>text(23,22,"x=y")
>rect(10,15,20,25)
>x1<-seq(10,15,0.1)
>segments(x1,15,x1,25)
>x2<-seq(15,20,0.1)
>segments(x2,x2,x2,25)
```

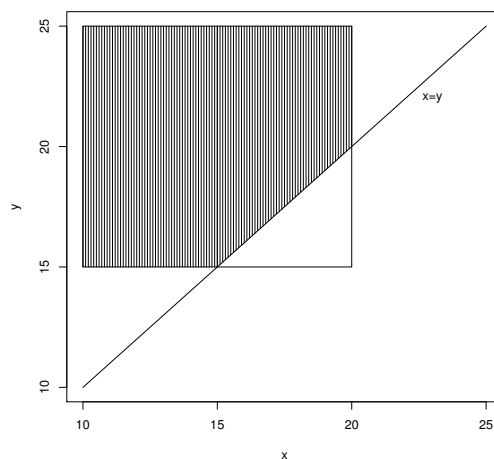


Figura 7.4: Rappresentazione del dominio \mathcal{D} .

Pertanto, tale probabilità può essere calcolata moltiplicando la densità congiunta per l'area tratteggiata rappresentata in Figura 7.4:

$$P(Y > X) = \frac{1}{100} \left[100 - \frac{25}{2} \right] = \frac{7}{8} = 0.875.$$

Calcoliamo ora tale probabilità utilizzando la simulazione delle variabili aleatorie A e B effettuando 1000000 simulazioni.

```
> N<-1000000
> x<-runif(N,min=10,max=20)
> y<-runif(N,min=15,max=25)
> diff<-y-x
> sum(diff>0)/length(diff)
[1] 0.875204
```

I due vettori x e y contengono 1000000 simulazioni dei tempi impiegati per fare la spesa dai due clienti A e B. Confrontando elemento per elemento i due vettori, costruiamo un nuovo vettore $\text{diff} = y - x$ contenente le differenze. Affinché B completi la spesa dopo di A occorre considerare gli elementi del vettore diff positivi. Per ottenere la probabilità che B completi la spesa dopo di A occorre considerare il rapporto tra i casi favorevoli e i casi possibili. I casi favorevoli sono il numero di elementi del vettore tali che $\text{diff} > 0$ e i casi possibili sono il numero di elementi dei vettori.

7.3 Distribuzione esponenziale

Nel caso discreto abbiamo mostrato che una variabile aleatoria caratterizzata da funzione di probabilità geometrica può utilizzarsi per descrivere il tempo di attesa per l'occorrenza del primo successo in una successione di prove ripetute di Bernoulli. Introduciamo ora la densità di probabilità esponenziale; questa può interpretarsi come l'analogo nel continuo della funzione di probabilità geometrica nel senso che una variabile aleatoria caratterizzata da densità di probabilità esponenziale può immaginarsi idonea a descrivere anch'essa un tempo di attesa che, però, in questo caso viene riguardato come variabile nel continuo.

Definizione 7.2 Sia $\lambda > 0$. Una variabile aleatoria X di funzione di distribuzione

$$F_X(x) = P(X \leq x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \geq 0 \end{cases} \quad (7.4)$$

e corrispondente densità di probabilità

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{altrimenti} \end{cases} \quad (7.5)$$

si dice *esponenzialmente distribuita con parametro λ* .

Nel seguito la notazione $X \sim \mathcal{E}(1, \lambda)$ verrà utilizzata per indicare che X ha distribuzione esponenziale di parametro λ ; X sarà anche detta *variabile esponenziale*. Per una variabile aleatoria esponenziale si ha

$$E(X) = \frac{1}{\lambda}, \quad E(X^2) = 2 \left(\frac{1}{\lambda}\right)^2, \quad \text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{1}{\lambda^2}.$$

Proposizione 7.1 Sia X una variabile aleatoria di densità (7.5). Per ogni s, t reali positivi risulta:

$$P(X > s + t | X > s) = P(X > t). \quad (7.6)$$

Se si interpreta X come un tempo di attesa, la (7.6) mostra che la probabilità condizionata che il tempo di attesa X sia maggiore di $t + s$ dato che essa è maggiore di s non dipende da quanto si è già atteso, ossia da s . Pertanto, la

distribuzione esponenziale, così come la distribuzione geometrica, gode della proprietà di “assenza di memoria”.

La funzione di distribuzione esponenziale riveste notevole importanza sia teorica che applicativa. Essa, ad esempio, interviene spesso quando si studiano sistemi di servizio in cui è ragionevole assumere che i tempi di interarrivo degli utenti oppure i tempi di espletamento dei servizi siano distribuiti proprio esponenzialmente, o allorché si considera la durata di funzionamento, realisticamente supposta aleatoria, di componenti elettronici o di dispositivi di varia natura.

Proposizione 7.2 *Sia X una variabile aleatoria di densità (7.5) e sia $Z = X - \tau$, con $\tau > 0$. Per ogni $z > 0$ risulta:*

$$P(Z \leq z | X > \tau) = P(X \leq z). \quad (7.7)$$

Nella Proposizione 7.2 interpretiamo ora X come durata di funzionamento di un componente elettronico, meccanico o di altra natura (ossia l'intervallo di tempo in cui esso funziona perfettamente) e, conseguentemente, $Z = X - \tau$ come durata della sua vita residua sapendo che esso ha già funzionato per una durata τ . La (7.7) mostra che la durata di vita residua ha la stessa distribuzione della durata di vita del componente considerato. Questa proprietà è un'ovvia conseguenza dell'assenza di memoria della funzione di distribuzione esponenziale. Ciò costituisce evidente circostanza che variabili aleatorie esponenzialmente distribuite non sono idonee a descrivere la durata di vita di dispositivi soggetti ad usura se non entro prefissati limiti di approssimazione, ad esempio nel caso di semplici dispositivi non soggetti a rapidi deterioramenti significativi, ma che possono subire danni per casi accidentali quali, ad esempio, corti circuiti, fulmini, imprevedibili sollecitazioni meccaniche, ecc.

R permette di calcolare la densità di probabilità, la funzione di distribuzione e i quantili di una variabile aleatoria esponenziale e anche di simulare tale variabile.

Si può richiedere ad R di eseguire direttamente il calcolo della densità esponenziale utilizzando la funzione

```
dexp(x, rate = lambda)
```

Gli argomenti di tale funzione sono

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria esponenziale;
- $rate$ è la frequenza λ della densità esponenziale.

Per calcolare la funzione di distribuzione invece utilizziamo la funzione

```
pexp(x, rate = lambda, lower.tail = TRUE)
```

Gli argomenti di tale funzione sono

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria esponenziale;

- *rate* è la frequenza λ della densità esponenziale;
- *lower.tail* se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

Ad esempio per rappresentare la densità di probabilità e la funzione di distribuzione di una variabile aleatoria esponenziale di valore medio $1/2$ utilizziamo il codice

```
>par(mfrow=c(1,2))
>curve(dexp(x,rate=2),from=0, to=10,xlab="x",
+ylab="f(x)",main="lambda=2")
>
>curve(pexp(x,rate=2),from=-2, to=10,
+xlab="x",ylab=expression(P(X<=x)),main="lambda=2")
```

che produce il grafico di Figura 7.5.

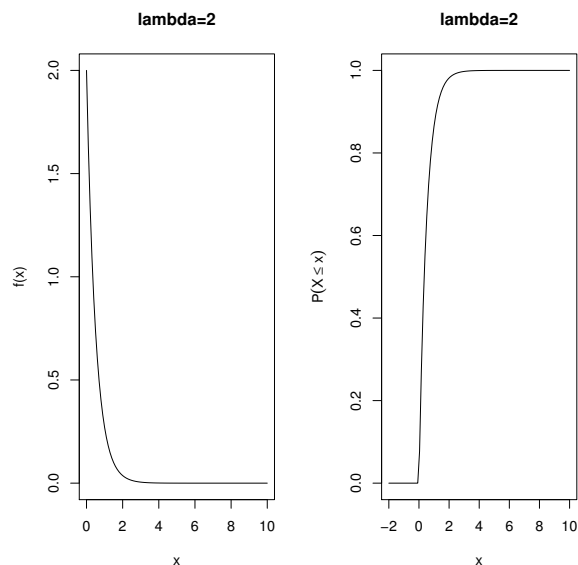


Figura 7.5: Rappresentazione della densità e della funzione di distribuzione esponenziale di valore medio $1/2$.

La probabilità che la variabile aleatoria esponenziale di valore medio $1/2$ assuma valori nell'intervallo $(0.5, 1.5)$ è

$$P(0.5 < X < 1.5) = P(X < 1.5) - P(X < 0.5) = e^{-2 \cdot 0.5} - e^{-2 \cdot 1.5} = 0.3180924$$

e corrisponde all'area sottesa dalla densità esponenziale in Figura 7.6 ottenuta tramite il seguente codice:


```
>curve(dexp(x,rate=2),from=0,to=2.5,xlab="x",ylab="f(x)")
>x<-seq(0.5,1.5,0.01)
>lines(x,dexp(x,rate=2),type="h",col="grey")
>text(1.1,0.5,"P(0.5<X<1.5)")
```

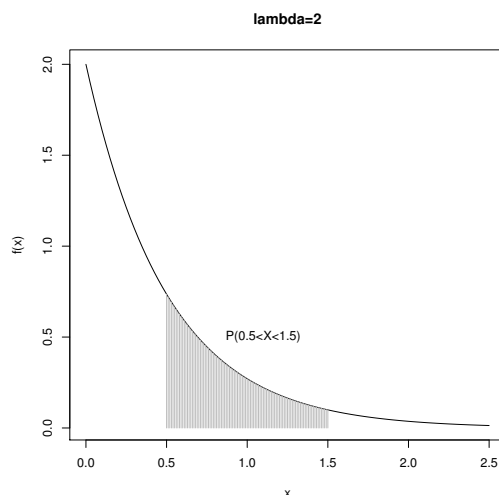


Figura 7.6: Rappresentazione della $P(0.5 < X < 1.5)$ per una variabile aleatoria esponenziale di valore medio $1/2$.

Come si evince dal grafico in Figura 7.6 la probabilità $P(0.5 < X < 1.5)$ può essere così valutata in R:

```
> pexp(1.5,2)-pexp(0.5,2)
[1] 0.3180924
```

In R si possono calcolare anche i quantili (percentili) della distribuzione esponenziale attraverso la funzione

```
qexp(z, rate = 1, lower.tail = TRUE)
```

dove

- z è il valore assunto (o i valori assunti) dalle probabilità relative al percentile $z \cdot 100$ -esimo;
- $rate$ è la frequenza λ della densità esponenziale;
- `lower.tail` se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

Il risultato della funzione è il percentile $z \cdot 100$ -esimo, ossia il più piccolo numero x assunto dalla variabile aleatoria esponenziale X tale che sussista la (7.3),

ossia che $P(X \leq x) \geq z$. Ad esempio, se si considera una variabile esponenziale di valore medio $1/2$, le seguenti linee di codice forniscono i quartili Q_0, Q_1, Q_2, Q_3, Q_4

```
>z<-c(0,0.25,0.5,0.75,1)
>qexp(z, rate=2)
[1] 0.0000000 0.1438410 0.3465736 0.6931472      Inf
```

che mostra che il primo quartile (25-esimo percentile) è $Q_1 = 0.1438410$, il secondo quartile o mediana (50-esimo percentile) è $Q_2 = 0.3465736$ e il terzo quartile (75-esimo percentile) è $Q_3 = 0.6931472$. Il minimo è $Q_0 = 0$ e il massimo è $Q_4 = \infty$. Infatti, ricordando che $P(X \leq x) = 1 - e^{-\lambda x}$ si ha che

$$1 - e^{-\lambda x} \geq z \iff x \geq -\frac{\ln(1-z)}{\lambda}$$

$$P(X \leq x) = 1 - e^{-2x} \geq 0.25 \iff x \geq -\frac{\ln(1-0.25)}{\lambda} = 0.1438410,$$

$$P(X \leq x) = 1 - e^{-2x} \geq 0.50 \iff x \geq -\frac{\ln(1-0.50)}{\lambda} = 0.3465736,$$

$$P(X \leq x) = 1 - e^{-2x} \geq 0.75 \iff x \geq -\frac{\ln(1-0.75)}{\lambda} = 0.6931472.$$

È possibile simulare in R la variabile aleatoria esponenziale generando una sequenza di numeri pseudocasuali mediante la funzione

```
rexp(N, rate=lambda)
```

dove

- N è lunghezza della sequenza da generare;
- $rate$ è la frequenza λ della densità esponenziale.

Il codice seguente

```
>par(mfrow=c(2,2))
>curve(dexp(x,rate=2),from=0, to=8,xlab="x",ylab="f(x)",
+ylim=c(0,2),main="Densità 'esponenziale', lambda=2")
>
>sim1<-rexp(500,rate=2)
>hist(sim1,freq=F,xlim=c(0,8),ylim=c(0,2),breaks=100,xlab="x",
+ylab="Istogramma",main="Densità 'simulata', N=500")
>
>sim2<-rexp(5000,rate=2)
>hist(sim2,freq=F,xlim=c(0,8),ylim=c(0,2),breaks=100,xlab="x",
+ylab="Istogramma",main="Densità 'simulata', N=5000")
>
>sim3<-rexp(50000,rate=2)
>hist(sim3,freq=F,xlim=c(0,8),ylim=c(0,2),breaks=100,xlab="x",
+ylab="Istogramma",main="Densità 'simulata', N=50000")
```

permette di confrontare in Figura 7.7 la densità esponenziale teorica di valore medio $1/2$ con la densità simulata scegliendo $N = 500, 5000, 50000$.

Si nota che all'aumentare del numero di simulazioni l'istogramma delle frequenze relative si avvicina alla densità esponenziale teorica.

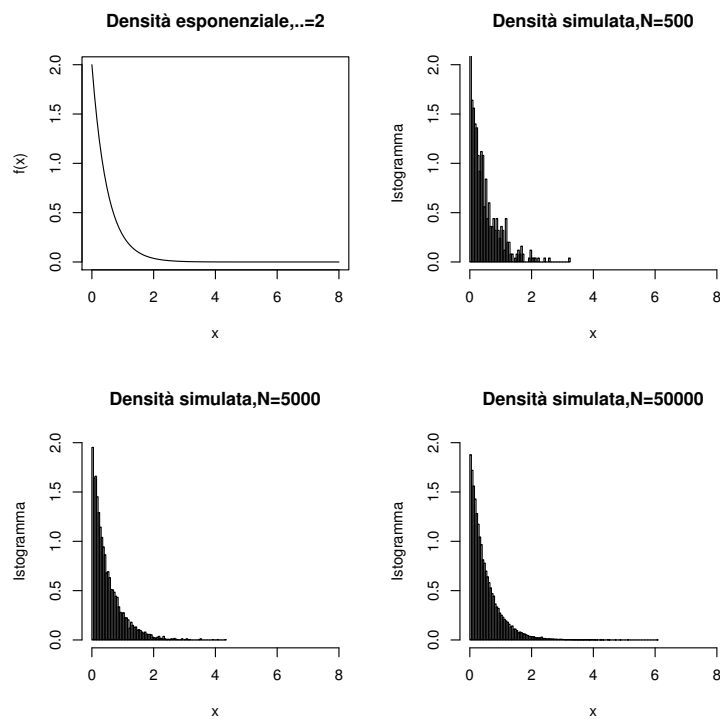


Figura 7.7: Confronto della densità esponenziale di valore medio $1/2$ con la densità simulata.

7.4 Distribuzione normale

La funzione di distribuzione normale, detta anche di Gauss o gaussiana, riveste estrema importanza nel calcolo delle probabilità e nella statistica anche in quanto essa costituisce una distribuzione limite alla quale tendono varie altre funzioni di distribuzioni sotto opportune ipotesi.

Definizione 7.3 Una variabile aleatoria X di densità di probabilità

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R} \quad (\mu \in \mathbb{R}, \sigma > 0), \quad (7.8)$$

si dice avere distribuzione normale di parametri μ e σ .

Dalla (7.8) si evince che per ogni $x \in \mathbb{R}$ risulta $f_X(\mu-x) = f_X(\mu+x)$; pertanto la densità normale è simmetrica rispetto all'asse $x = \mu$. La densità $f_X(x)$ presenta il massimo $(\sigma\sqrt{2\pi})^{-1}$ nel punto di ascissa $x = \mu$ e due flessi nei punti di ascisse $\mu - \sigma$ e $\mu + \sigma$. Il grafico di $f_X(x)$ esibisce una caratteristica forma a campana, simmetrica rispetto a $x = \mu$. La notazione $X \sim \mathcal{N}(\mu, \sigma)$ verrà utilizzata nel seguito per indicare che X ha distribuzione normale di parametri μ e σ , o più semplicemente che è una *variabile normale*.

In R la densità normale si calcola attraverso la funzione

```
dnorm(x, mean = mu, sd = sigma)
```

dove

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria normale;
- mean e sd sono il valore medio e la deviazione standard della densità normale.

Il seguente codice permette di visualizzare la densità di $X \sim \mathcal{N}(\mu, 1)$ con $\mu = -3, -2, -1, 0, 1, 2, 3$.

```
>curve(dnorm(x,mean=-3,sd=1),from=-6, to=6,xlab="x",ylab="f(x)",
+main="mu=-3,-2,-1,0,1,2,3;sigma=1")
>curve(dnorm(x,mean=-2,sd=1),from=-6, to=6,xlab="x",ylab="f(x)",
+add=TRUE)
>curve(dnorm(x,mean=-1,sd=1),from=-6, to=6,xlab="x",ylab="f(x)",
+add=TRUE)
>curve(dnorm(x,mean=0,sd=1),from=-6, to=6,xlab="x",ylab="f(x)",
+add=TRUE,lty=2)
>curve(dnorm(x,mean=1,sd=1),from=-6, to=6,xlab="x",ylab="f(x)",
+add=TRUE)
>curve(dnorm(x,mean=2,sd=1),from=-6, to=6,xlab="x",ylab="f(x)",
+add=TRUE)
>curve(dnorm(x,mean=3,sd=1),from=-6, to=6,xlab="x",ylab="f(x)",
+add=TRUE)
```

Come illustrato nella Figura 7.8 variazioni del parametro μ comportano traslazioni della curva lungo l'asse delle ascisse; infatti, al crescere del parametro μ la curva si sposta lungo l'asse delle ascisse senza cambiare forma.

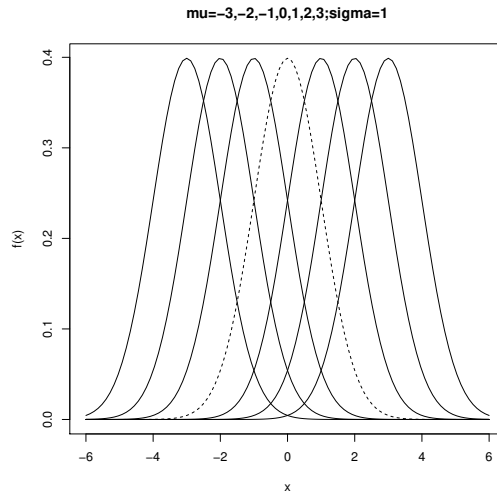


Figura 7.8: Densità normale al variare di $\mu = -3, -2, -1, 0, 1, 2, 3$ (da sinistra verso destra).

Il parametro σ , pari alla semiampiezza tra i due punti di flesso, caratterizza la larghezza della funzione. Poiché l'ordinata massima è inversamente proporzionale a σ , al crescere di σ questa decresce, mentre l'area sottesa dalla densità deve rimanere unitaria. Il seguente codice permette di visualizzare la densità di $X \sim \mathcal{N}(0, \sigma)$ con $\sigma = 0.5, 1, 1.5$.

```
>curve(dnorm(x,mean=0,sd=0.5),from=-4, to=4,xlab="x",
+ylab="f(x)",main="mu=0; sigma=0.5,1,1.5")
>curve(dnorm(x,mean=0,sd=1),from=-4, to=4,xlab="x",
+ylab="f(x)",add=TRUE,lty=2)
>curve(dnorm(x,mean=0,sd=1.5),from=-4, to=4,xlab="x",
+ylab="f(x)",add=TRUE)
```

il cui grafico è riportato in Figura 7.9. Si nota che al crescere di σ la curva diventa sempre più piatta, mentre al decrescere di σ essa si allunga verso l'alto restringendosi contemporaneamente ai lati.

La funzione di distribuzione di una variabile aleatoria $X \sim \mathcal{N}(\mu, \sigma)$ è:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(y) dy = \Phi\left(\frac{x-\mu}{\sigma}\right) \quad x \in \mathbb{R} \quad (7.9)$$

dove

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z \exp\left\{-\frac{y^2}{2}\right\} dy, \quad z \in \mathbb{R}. \quad (7.10)$$

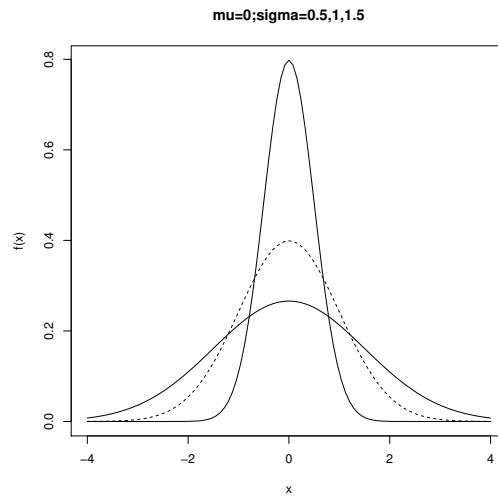


Figura 7.9: Densità normale al variare di $\sigma = 0.5, 1, 1.5$ (dall'alto verso il basso in prossimità dell'origine).

è la funzione di distribuzione di una variabile aleatoria $Z \sim \mathcal{N}(0, 1)$, detta *normale standard*. Pertanto, se $X \sim \mathcal{N}(\mu, \sigma)$ si ha:

$$P(a < X < b) = F_X(b) - F_X(a) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right). \quad (7.11)$$

In R la funzione di distribuzione di una variabile $X \sim \mathcal{N}(\mu, \sigma)$ si calcola tramite la funzione:

```
>pnorm(x, mean = mu, sd = sigma, lower.tail = TRUE)
```

dove:

- `x` è il valore assunto (o i valori assunti) dalla variabile aleatoria normale;
- `mean` e `sd` sono il valore medio e la deviazione standard della densità normale;
- `lower.tail` se tale parametro è `TRUE` (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è `FALSE` calcola $P(X > x)$.

Il seguente codice permette di visualizzare la funzione di distribuzione di $X \sim \mathcal{N}(0, \sigma)$ con $\sigma = 0.5, 1, 1.5$:

```
>curve(pnorm(x, mean=0, sd=0.5), from=-4, to=4, xlab="x",
+ylab=expression(P(X<=x)), main="mu=0; sigma=0.5, 1, 1.5", lty=2)
>text(-0.4, 0.8, "sigma=0.5")
>curve(pnorm(x, mean=0, sd=1), add=TRUE)
>arrows(-1, 0.1, 0.5, 0.2, code=1, length = 0.10)
>text(0.8, 0.2, "sigma=1")
>curve(pnorm(x, mean=0, sd=1.5), add=TRUE, lty=3)
>text(-2.2, 0.2, "sigma=1.5")
```

il cui grafico è riportato in Figura 7.10. La funzione `arrows()` ha come argomenti

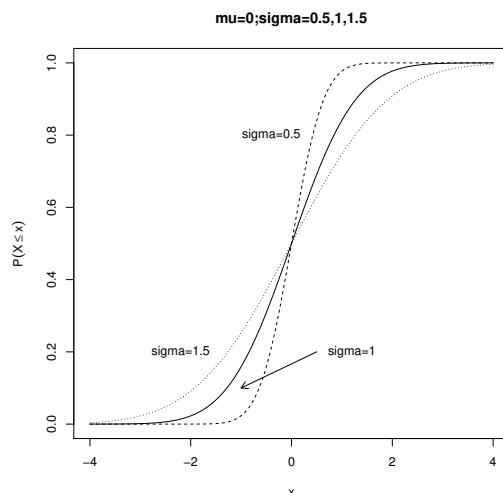


Figura 7.10: Funzione di distribuzione normale al variare di $\sigma = 0.5, 1, 1.5$.

le due coordinate della linea della freccia, il parametro `code` può assumere i valori 1,2,3 a seconda se la freccia deve essere unidirezionale verso sinistra, unidirezionale verso destra oppure bidirezionale; il parametro `length` fornisce invece la grandezza della freccia.

⇒ Regola del 3σ

Per una qualsiasi variabile aleatoria normale $X \sim \mathcal{N}(\mu, \sigma)$ risulta

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = P\left(-3 < \frac{X - \mu}{\sigma} < 3\right) = P(-3 < Z < 3) = 0.9973002.$$

Quindi la probabilità che una variabile aleatoria $X \sim \mathcal{N}(\mu, \sigma)$ assuma valori in un intervallo avente come centro μ e semiampiezza 3σ è prossima all'unità. Questa proprietà delle variabili aleatorie normali è nota come *regola del 3σ* . Infatti, utilizzando R, per una variabile aleatoria normale $Z \sim \mathcal{N}(0, 1)$ si ha

```
pnorm(3, mean=0, sd=1) - pnorm(-3, mean=0, sd=1)
[1] 0.9973002
```

La regola del 3σ permette di individuare l'intervallo $(\mu - 3\sigma, \mu + 3\sigma)$ in cui rappresentare la funzione densità di una variabile normale di valore medio μ e varianza σ^2 in maniera tale che l'area sottesa dalla curva sia circa unitaria e l'area delle code destra e sinistra sia trascurabile.

In R si possono calcolare anche i quantili (percentili) della distribuzione normale attraverso la funzione

```
qnorm(z, mean = mu, sd = sigma, lower.tail = TRUE)
```

dove

- z è il valore assunto (o i valori assunti) dalle probabilità relative al percentile $z \cdot 100$ -esimo;
- mean e sd sono il valore medio e la deviazione standard della densità normale;
- lower.tail se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

Il risultato della funzione è il percentile $z \cdot 100$ -esimo, ossia il più piccolo numero x assunto dalla variabile aleatoria normale X tale che sussista la (7.3), ossia che $P(X \leq x) \geq z$. Ad esempio, se si considera una variabile normale standard $Z \sim \mathcal{N}(0, 1)$, le seguenti linee di codice forniscono i quartili Q_0, Q_1, Q_2, Q_3, Q_4

```
>z<-c(0,0.25,0.5,0.75,1)
>qnorm(z, mean = 0, sd = 1)
[1] -Inf -0.6744898 0.0000000 0.6744898 Inf
```

che mostra che il primo quartile (25-esimo percentile) è $Q_1 = -0.6744898$, il secondo quartile o mediana (50-esimo percentile) è $Q_2 = 0$ e il terzo quartile (75-esimo percentile) è $Q_3 = 0.6744898$ (per la simmetria intorno all'origine della densità normale standard). Il minimo è $Q_0 = -\infty$ e il massimo è $Q_4 = \infty$.

Esempio 7.2 Sia $X \sim \mathcal{N}(\mu, \sigma)$, con $\mu \in \mathbb{R}$ e $\sigma > 0$. Determinare il reale ε tale che $P(X - \mu \leq \varepsilon) = 0.975$.

Osserviamo che

$$P(X - \mu \leq \varepsilon) = P\left(\frac{X - \mu}{\sigma} \leq \frac{\varepsilon}{\sigma}\right) = P\left(Z \leq \frac{\varepsilon}{\sigma}\right) = 0.975$$

dove Z è una variabile aleatoria normale standard. Per determinare il valore da assegnare a ε/σ si può utilizzare in R la funzione quantile, ottenendo:

```
> qnorm(0.975, mean=0, sd=1)
[1] 1.959964
```

da cui segue che $\varepsilon = 1.959964 \cdot \sigma$. \diamond

Esempio 7.3 Supponiamo che $X \sim \mathcal{N}(\mu, \sigma)$, con $\mu = 0$ e $\sigma = 0.01$, descriva l'errore di misura nel valutare una certa distanza. Determiniamo la probabilità p che il valore assoluto dell'errore di misura sia minore di $\beta = 0.02$.

Occorre quindi determinare $p = P(|X| < 0.02) = P(-0.02 < X < 0.02)$. Facendo uso di R si ha

```
> pnorm(0.02, mean=0, sd=0.01) - pnorm(-0.02, mean=0, sd=0.01)
[1] 0.9544997
```

che mostra che la probabilità richiesta è $p = 0.9544997$. \diamond

È possibile simulare in R la variabile aleatoria normale generando una sequenza di numeri pseudocasuali mediante la funzione

```
> rnorm(N, mean = mu, sd = sigma)
```

dove:

- N è la lunghezza della sequenza da generare;
- mean e sd sono il valore medio e la deviazione standard della densità normale;

Il codice seguente

```
> par(mfrow=c(2,2))
> curve(dnorm(x, mean=2, sd=1), from=-2, to=6, xlab="x", ylab="f(x)",
+ ylim=c(0,0.5), main="Densità 'normale, mu=2, sigma=1")
>
> sim1<-rnorm(500, mean=2, sd=1)
> hist(sim1, freq=F, xlim=c(-2,6), ylim=c(0,0.5), breaks=100, xlab="x",
+ ylab="Istogramma", main="Densità 'simulata, N=500")
>
> sim2<-rnorm(5000, mean=2, sd=1)
> hist(sim2, freq=F, xlim=c(-2,6), ylim=c(0,0.5), breaks=100, xlab="x",
+ ylab="Istogramma", main="Densità 'simulata, N=5000")
>
> sim3<-rnorm(50000, mean=2, sd=1)
> hist(sim3, freq=F, xlim=c(-2,6), ylim=c(0,0.5), breaks=100, xlab="x",
+ ylab="Istogramma", main="Densità 'simulata, N=50000")
```

permette di confrontare in Figura 7.11 la densità normale teorica con $\mu = 2$, $\sigma = 1$ con la densità simulata scegliendo $N = 500, 5000, 50000$. All'aumentare del numero di simulazioni l'istogramma delle frequenze relative si avvicina sempre di più alla densità esponenziale teorica.

Osserviamo infine che se X_1, X_2, \dots, X_n sono variabili aleatorie indipendenti con $X_i \sim \mathcal{N}(\mu_i, \sigma_i)$ per $i = 1, 2, \dots, n$ e se a_1, a_2, \dots, a_n sono numeri reali, allora

$$Y = a_1 X_1 + a_2 X_2 + \dots + a_n X_n$$

ha distribuzione normale di parametri μ e σ , dove

$$\begin{aligned}\mu &= E(Y) = a_1 \mu_1 + a_2 \mu_2 + \dots + a_n \mu_n \\ \sigma^2 &= \text{Var}(Y) = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \dots + a_n^2 \sigma_n^2.\end{aligned}$$

⇒ Approssimazione della distribuzione binomiale con la distribuzione normale

Il calcolo delle probabilità binomiali diviene rapidamente oneroso al crescere di n . È quindi utile ricercare delle formule approssimate in grado di rendere agevole tale calcolo e, al contempo, accettabile l'errore derivante dall'approssimazione. Prenderemo in primo luogo in considerazione il *teorema di De Moivre-Laplace* e successivamente il *teorema centrale di convergenza*.

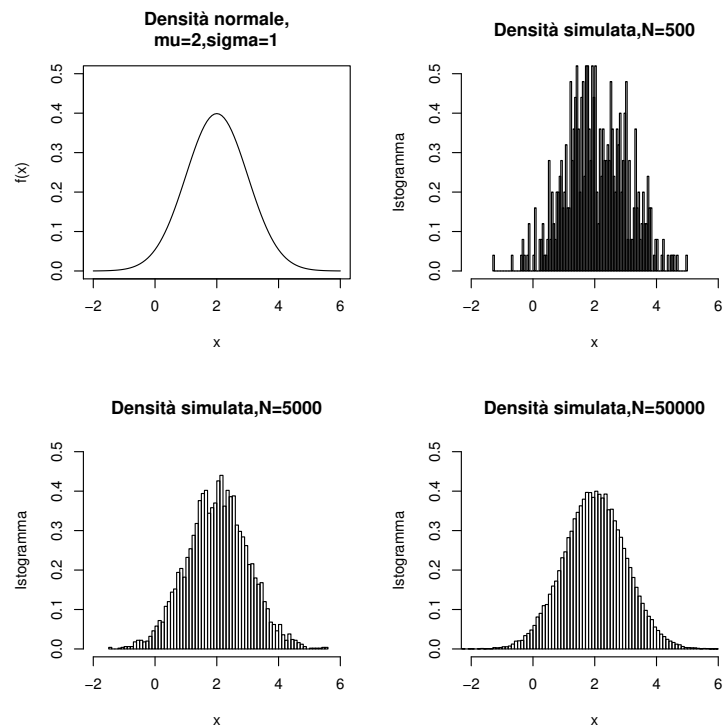


Figura 7.11: Confronto della densità normale con $\mu = 2$ e $\sigma = 1$ con la densità simulata.

Teorema 7.1 (Teorema di De Moivre-Laplace) Sia X_1, X_2, \dots una successione di variabili aleatorie indipendenti distribuite alla Bernoulli con parametro p ($0 < p < 1$), e sia $Y_n = X_1 + X_2 + \dots + X_n$. Allora per ogni $x \in \mathbb{R}$ risulta:

$$\lim_{n \rightarrow +\infty} P\left(\frac{Y_n - np}{\sqrt{np(1-p)}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy,$$

ossia

$$\frac{Y_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} Z,$$

converge in distribuzione alla variabile aleatoria Z normale standard.

Ricordiamo che se X_1, X_2, \dots sono variabili aleatorie indipendenti di Bernoulli di parametro p , allora $Y_n = X_1 + X_2 + \dots + X_n$ è una variabile aleatoria binomiale di valore medio np e varianza $np(1-p)$. Il Teorema 7.1 mostra che sottraendo a Y_n la sua media np e dividendo la differenza per la deviazione standard $\sqrt{np(1-p)}$, si ottiene una variabile aleatoria standardizzata la cui funzione di distribuzione è per n grande approssimativamente normale standard. La bontà dell'approssimazione dipende da n e da p e migliora al tendere di p a $1/2$. In generale si suole assumere che l'approssimazione sia soddisfacente per $n > 10$ e per $5/n < p < 1 - 5/n$.

Esaminiamo ora l'approssimazione della binomiale alla normale

$$Y_n \simeq np + \sqrt{np(1-p)} Z, \quad (7.12)$$

al variare di n con p fissato. Si noti che il secondo membro della (7.12) è una variabile aleatoria con densità normale di valore medio np e varianza $np(1-p)$. Il codice seguente confronta la densità normale di valore medio np e varianza $np(1-p)$ e la funzione di probabilità binomiale per $n = 25, 50, 75, 100$ e $p = 0.2$

```
>par(mfrow=c(2,2))
>p<-0.2
>q<-1-p
>x<-0:25
>n<-25
>curve(dnorm(x,np,sqrt(np*q)),from=np-3*sqrt(np*q),
+to=np+3*sqrt(np*q),xlab="x",ylab="P(X=x)",
+main="Binomiale,n=25,p=0.2")
>lines(x,dbinom(x,n,0.2),type="h")
>
>x<-0:50
>n<-50
>curve(dnorm(x,np,sqrt(np*q)),from=np-3*sqrt(np*q),
+to=np+3*sqrt(np*q),xlab="x",ylab="P(X=x)",
+main="Binomiale,n=50,p=0.2")
>lines(x,dbinom(x,n,0.2),type="h")
>
>x<-0:75
>n<-75
>curve(dnorm(x,np,sqrt(np*q)),from=np-3*sqrt(np*q),
+to=np+3*sqrt(np*q),xlab="x",ylab="P(X=x)",
+main="Binomiale,n=75,p=0.2")
```

```

>lines(x,dbinom(x,n,0.2),type="h")
>
>x<-0:100
>n<-100
>curve(dnorm(x,n*p,sqrt(n*p*q)),from=n*p-3*sqrt(n*p*q),
+to=n*p+3*sqrt(n*p*q),xlab="x",ylab="P(X=x)",
+main="Binomiale,n=100,p=0.2")
>lines(x,dbinom(x,n,0.2),type="h")

```

il cui grafico è riportato in Figura 7.12. Si nota che l'approssimazione migliora al crescere di n .

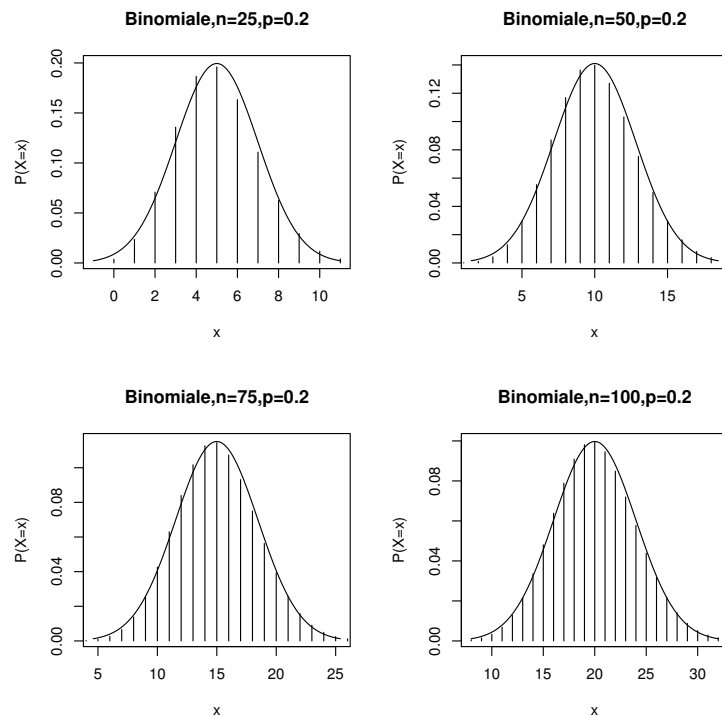


Figura 7.12: Confronto della probabilità binomiale della variabile $Y_n \sim \mathcal{B}(n, 0.2)$ con la densità normale di valor medio $\mu = np$ e deviazione standard $\sigma = \sqrt{np(1-p)}$ per varie scelte di n .

Esaminiamo ora l'approssimazione (7.12) della binomiale alla normale al variare di p con n fissato. Il codice seguente confronta la densità normale di valore medio np e varianza $np(1-p)$ e la funzione di probabilità binomiale per $n = 20$ e $p = 0.125, 0.25, 0.375, 0.5$

```

>par(mfrow=c(2,2))
>x<-0:20
>n<-20

```

```

>p<-0.125
>q<-1-p
>curve(dnorm(x,n*p,sqrt(n*p*q)),from=n*p-3*sqrt(n*p*q),
+to=n*p+3*sqrt(n*p*q),,xlab="x",ylab="P(X=x)",
+main="Binomiale,n=20,p=0.125")
>lines(x,dbinom(x,n,0.125),type="h")
>
>p<-0.25
>q<-1-p
>curve(dnorm(x,n*p,sqrt(n*p*q)),from=n*p-3*sqrt(n*p*q),
+to=n*p+3*sqrt(n*p*q),xlab="x",ylab="P(X=x)",
+main="Binomiale,n=20,p=0.25")
>lines(x,dbinom(x,n,0.25),type="h")
>
>p<-0.375
>q<-1-p
>curve(dnorm(x,n*p,sqrt(n*p*q)),from=n*p-3*sqrt(n*p*q),
+to=n*p+3*sqrt(n*p*q),xlab="x",ylab="P(X=x)",
+main="Binomiale,n=20,p=0.375")
>lines(x,dbinom(x,n,0.375),type="h")
>
>p<-0.5
>q<-1-p
>curve(dnorm(x,n*p,sqrt(n*p*q)),from=n*p-3*sqrt(n*p*q),
+to=n*p+3*sqrt(n*p*q),xlab="x",ylab="P(X=x)",
+main="Binomiale,n=20,p=0.5")
>lines(x,dbinom(x,n,0.5),type="h")

```

il cui grafico è riportato in Figura 7.13. Si nota che l'approssimazione non è buona per piccoli valori di p e migliora al tendere di p a $1/2$, diventando poi eccellente quando $p = 1/2$.

⇒ Teorema centrale di convergenza

Vogliamo ora introdurre uno dei più importanti risultati della teoria della probabilità, noto quale *teorema centrale di convergenza* o *teorema centrale del limite*, che fornisce una semplice ed utile approssimazione alla distribuzione della somma di variabili aleatorie indipendenti, evidenziando al contempo la grande importanza della distribuzione normale.

Teorema 7.2 (Teorema centrale di convergenza) *Sia X_1, X_2, \dots una successione di variabili aleatorie, definite nello stesso spazio di probabilità, indipendenti e identicamente distribuite con valore medio μ finito e varianza σ^2 finita e positiva. Posto per ogni intero n positivo $Y_n = X_1 + X_2 + \dots + X_n$, per ogni $x \in \mathbb{R}$ risulta:*

$$\lim_{n \rightarrow +\infty} P\left(\frac{Y_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy = \Phi(x), \quad (7.13)$$

ossia la successione delle variabili aleatorie standardizzate

$$\frac{Y_n - E(Y_n)}{\sqrt{\text{Var}(Y_n)}} = \frac{Y_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} Z,$$

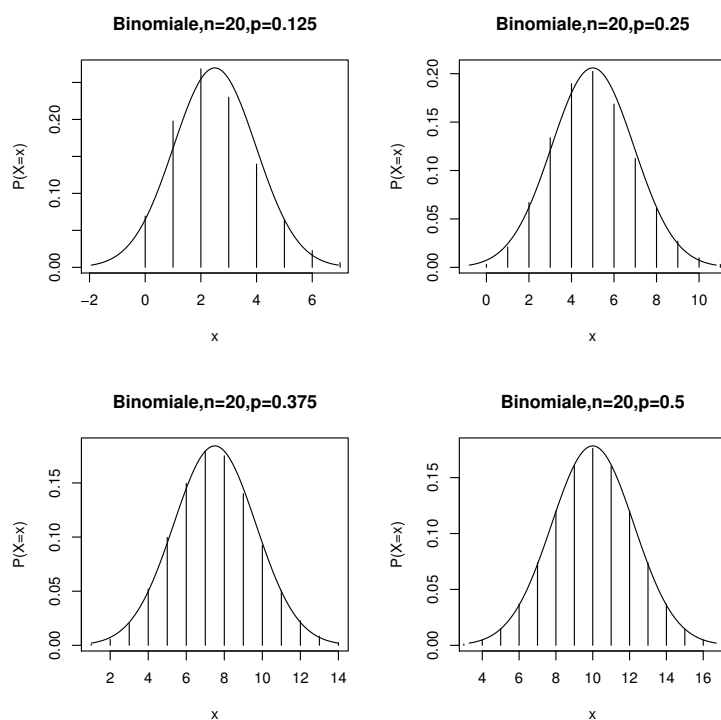


Figura 7.13: Confronto della probabilità binomiale della variabile $Y_n \sim \mathcal{B}(20, p)$ con la densità normale di valor medio $\mu = np$ e deviazione standard $\sigma = \sqrt{np(1-p)}$ per varie scelte di p .

converge in distribuzione alla variabile aleatoria normale standard.

Il Teorema 7.2 mostra inoltre che sottraendo a $X_1 + X_2 + \dots + X_n$ la sua media $n\mu$ e dividendo la differenza per la deviazione standard di Y_n , ossia per $\sigma\sqrt{n}$, si ottiene una variabile aleatoria standardizzata la cui funzione di distribuzione è per n sufficientemente grande approssimativamente normale standard. Quindi, per n grande la distribuzione di $Y_n = X_1 + X_2 + \dots + X_n$ è approssimativamente normale con valore medio $n\mu$ e varianza $n\sigma^2$. Va menzionato che la bontà delle approssimazioni dipende da n e dal tipo di distribuzione delle variabili X_1, X_2, \dots, X_n . L'approssimazione migliora al crescere di n e nelle applicazioni spesso si verifica che essa è già soddisfacente per $n \geq 30$. \Rightarrow

\Rightarrow Approssimazione della distribuzione di Poisson con la distribuzione normale

Se, ad esempio, supponiamo che X_1, X_2, \dots è una *successione di variabili aleatorie indipendenti di Poisson* di parametro λ allora $Y_n = X_1 + X_2 + \dots + X_n$ è ancora una variabile aleatoria di Poisson di parametro $n\lambda$. Quindi, il teorema centrale di convergenza afferma che per n grande la distribuzione di $Y_n = X_1 + X_2 + \dots + X_n$ è approssimativamente normale con valore medio $n\lambda$ e varianza $n\lambda$, ossia

$$Y_n \sim n\lambda + \sqrt{n\lambda}Z.$$

dove $n\lambda + \sqrt{n\lambda}Z$ è una variabile aleatoria con densità normale di valore medio $n\lambda$ e varianza $n\lambda$. Esaminiamo ora l'approssimazione della distribuzione di Poisson di parametro $n\lambda$ alla normale di valore medio e varianza $n\lambda$ al variare del parametro $n\lambda$. Il seguente codice

```
>par(mfrow=c(2,2))
>x<-0:100
>curve(dnorm(x,5,sqrt(5)),from=5-3*sqrt(5),to=5+3*sqrt(5),
+xlabs="x",ylab="P(X=x)",main="Poisson , n_lambd=5")
>lines(x,dpois(x, 5),type="h")
>
>curve(dnorm(x,10,sqrt(10)),from=10-3*sqrt(10),to=10+3*sqrt(10),
+xlabs="x",ylab="P(X=x)",main="Poisson , n_lambd=10")
>lines(x,dpois(x, 10),type="h")
>
>curve(dnorm(x,25,sqrt(25)),from=25-3*sqrt(25),to=25+3*sqrt(25),
+xlabs="x",ylab="P(X=x)",
+main="Poisson , n_lambd=25")
>lines(x,dpois(x, 25),type="h")
>
>curve(dnorm(x,50,sqrt(50)),from=50-3*sqrt(50),to=50+3*sqrt(50),
+xlabs="x",ylab="P(X=x)",main="Poisson , n_lambd=50")
>lines(x,dpois(x, 50),type="h")
```

permette di visualizzare la Figura 7.14 in cui si confronta la probabilità di Poisson di parametro $n\lambda$ con la densità normale di valore medio e varianza $n\lambda = 5, 10, 25, 50$. Si nota che al crescere di $n\lambda$ aumenta l'accuratezza dell'approssimazione.

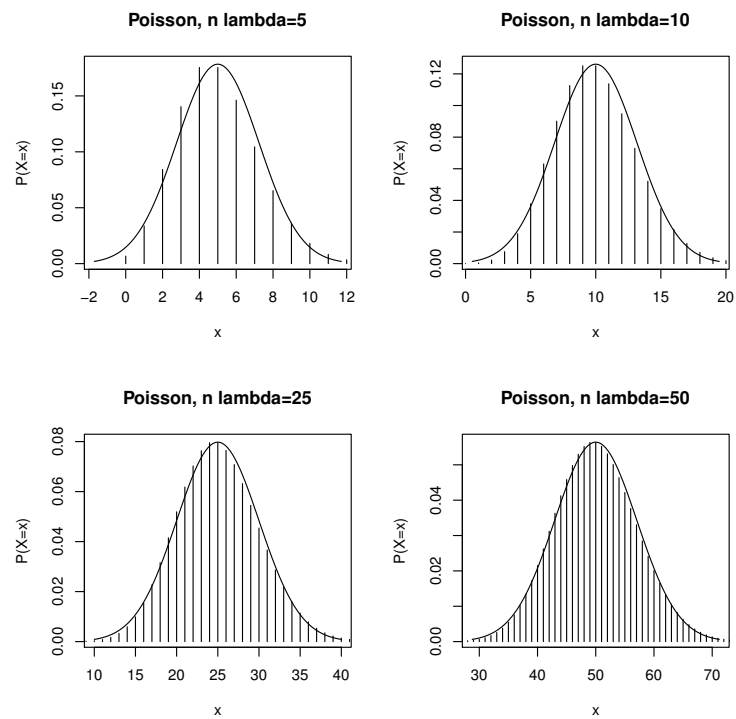


Figura 7.14: Confronto della probabilità di Poisson della variabile $X \sim \mathcal{P}(n\lambda)$ con la densità normale di valor medio $\mu = n\lambda$ e deviazione standard $\sigma = \sqrt{n\lambda}$ per varie scelte di $n\lambda$.

7.5 Distribuzione chi-quadrato

Per definire la densità chi-quadrato occorre introdurre prima la funzione $\Gamma(\nu)$, così definita:

$$\Gamma(\nu) = \int_0^{+\infty} x^{\nu-1} e^{-x} dx, \quad \nu > 0, \quad (7.14)$$

Se $\nu > 1$ per la funzione $\Gamma(\nu)$ sussiste la seguente proprietà di fattorizzazione:

$$\Gamma(\nu) = (\nu - 1) \Gamma(\nu - 1) \quad (\nu > 1), \quad (7.15)$$

La funzione gamma è una generalizzazione dei fattoriali; infatti, se ν è un intero positivo, usando iterativamente la (7.15), si ottiene:

$$\Gamma(\nu) = (\nu - 1)! \quad (\nu = 1, 2, \dots),$$

avendo fatto uso della proprietà $\Gamma(1) = 1$ direttamente ricavata dalla (7.14).

In R la funzione $\Gamma(\nu)$ si calcola semplicemente tramite la funzione `gamma`(ν)
Ad esempio, risulta:

```
> gamma(1/2)
[1] 1.772454
> gamma(3/2)
[1] 0.886227
```

che mostra che $\Gamma(1/2) = \sqrt{\pi}$ e $\Gamma(3/2) = (1/2)\Gamma(1/2) = \sqrt{\pi}/2$.

Possiamo ora definire la *densità chi-quadrato*.

Definizione 7.4 Una variabile aleatoria X di densità di probabilità

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(n/2)} \left(\frac{1}{2}\right)^{n/2} x^{n/2-1} e^{-x/2}, & x > 0 \\ 0, & x \leq 0 \end{cases} \quad (7.16)$$

con n intero positivo e con $\Gamma(\nu)$ definita in (7.14), si dice di *distribuzione chi-quadrato con n gradi di libertà*.

Nel seguito con $X \sim \chi^2(n)$ intenderemo che X ha distribuzione chi-quadrato con n gradi di libertà. In particolare, quando $n = 2$, (7.16) corrisponde ad una densità esponenziale di valore medio $1/\lambda = 2$.

Il seguente teorema evidenzia il ruolo giocato dal numero di gradi di libertà n

Teorema 7.3 Siano X_1, X_2, \dots, X_n variabili aleatorie indipendenti, con $X_i \sim \mathcal{N}(0, 1)$ per $i = 1, 2, \dots, n$. Allora, $Y_n = X_1^2 + X_2^2 + \dots + X_n^2$ ha distribuzione chi-quadrato con n gradi di libertà.

Il Teorema 7.3 afferma che la somma dei quadrati di variabili aleatorie normali standard indipendenti ha distribuzione chi-quadrato con un numero di gradi di libertà uguale al numero degli addendi. Quindi, la denominazione “numero di

gradi di libertà”, attribuita al parametro n , assume il significato di numero di addendi indipendenti presenti nella somma.

Il valore medio, momento del secondo ordine e la varianza di una variabile chi-quadrato con n gradi di libertà sono

$$E(X) = n, \quad E(X^2) = n(n+2), \quad \text{Var}(X) = 2n.$$

R permette di calcolare la densità di probabilità, la funzione di distribuzione e i quantili di una variabile aleatoria chi-quadrato e anche di simulare tale variabile.

Si può richiedere ad R di eseguire direttamente il calcolo della densità chi-quadrato utilizzando la funzione

```
dchisq(x, df)
```

Gli argomenti di tale funzione sono

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria chi-quadrato;
- df numero di gradi di libertà (non negativo, può anche essere un valore non intero).

Per calcolare la funzione di distribuzione invece utilizziamo la funzione

```
pchisq(x, df, lower.tail = TRUE)
```

Gli argomenti di tale funzione sono

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria chi-quadrato;
- df numero di gradi di libertà;
- `lower.tail` se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

Il codice seguente

```
>par(mfrow=c(1,2))
>curve(dchisq(x,df=1),from=0, to=18,ylim=c(0,0.3),xlab="x",
+ylab="f(x)",main="n=1,3,5,7")
>text(3,0.27,"n=1")
>curve(dchisq(x,df=3),add=TRUE,lty=2)
>text(4,0.20,"n=3")
>
>curve(dchisq(x,df=5),add=TRUE,lty=3)
>text(6,0.14,"n=5")
>curve(dchisq(x,df=7),add=TRUE,lty=4)
>text(11,0.08,"n=7")
>
>curve(pchisq(x,df=1),from=-2, to=18,ylim=c(0,1),xlab="x",
+ylab=expression(P(X<=x)),main="n=1,3,5,7")
>text(0,0.9,"n=1")
>
>curve(pchisq(x,df=3),add=TRUE,lty=2)
>arrows(4,0.7,12,0.8,code=1,length = 0.10)
>text(14,0.8,"n=3")
```

```

>
>curve(pchisq(x,df=5),add=TRUE,lty=3)
>arrows(5.5,0.6,13,0.7,code=1,length = 0.10)
>text(15,0.7,"n=5")
>
>curve(pchisq(x,df=7),add=TRUE,lty=4)
>text(8,0.4,"n=7")

```

permette di rappresentare in Figura 7.15 la densità di probabilità e la funzione di distribuzione di una variabile aleatoria $X \sim \chi^2(n)$ per $n = 1, 3, 5, 7$. La funzione densità chi-quadrato con n gradi di libertà è strettamente decrescente per $n = 1, 2$, mentre per $n > 2$ presenta un unico punto di massimo in $x = n - 2$.

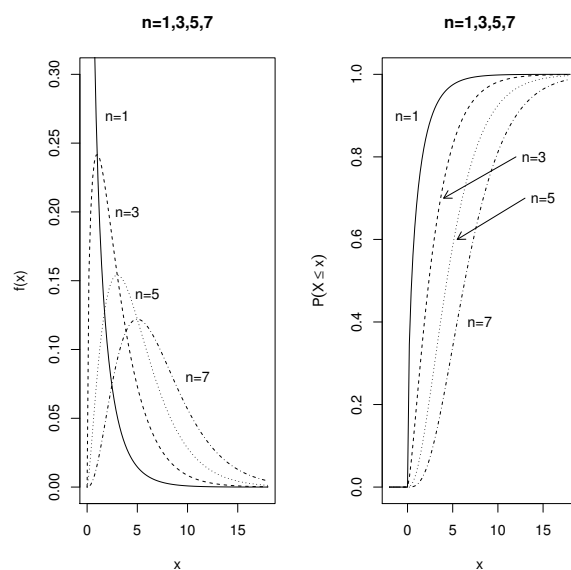


Figura 7.15: Densità di probabilità e funzione di distribuzione di $X \sim \chi^2(n)$.

È anche possibile calcolare i quantili e simulare una variabile chi-quadrato tramite le funzioni

```

qchisq(p, df, lower.tail = TRUE)
rchisq(N, df)

```

dove df indica il numero di gradi di libertà; gli altri parametri sono già stati descritti per le altre distribuzioni.

Osserviamo infine che se X_1, X_2, \dots, X_n sono variabili aleatorie indipendenti, con $X_i \sim \chi^2(k_i)$ per $i = 1, 2, \dots, n$, allora la variabile aleatoria

$$Y = X_1 + X_2 + \dots + X_n$$

è caratterizzata da densità chi-quadrato con $k = k_1 + k_2 + \dots + k_n$ gradi di libertà.

La distribuzione chi-quadrato riveste un ruolo importante in statistica nella stima puntuale e intervallare della varianza di una popolazione normale, ed anche in molti test di verifica di ipotesi statistiche.

7.6 Distribuzione di Student

Un'altra distribuzione di considerevole interesse applicativo è quella di Student¹ che passiamo a definire.

Definizione 7.5 Una variabile aleatoria X di densità di probabilità

$$f_X(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi} \Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}, \quad x \in \mathbb{R} \quad (7.17)$$

con n intero positivo e con $\Gamma(\nu)$ definita in (7.14), si dice avere distribuzione di Student, o avere “distribuzione t di Student”, con n gradi di libertà.

Nel seguito con $X \sim \mathcal{T}(n)$ intenderemo che X ha distribuzione di Student con n gradi di libertà. Il teorema seguente mostra la stretta connessione esistente tra una variabile a distribuzione di Student (7.17) e variabili a distribuzione chi-quadrato e normale standard.

Teorema 7.4 Siano $Y \sim \chi^2(n)$ e $Z \sim \mathcal{N}(0, 1)$ variabili aleatorie indipendenti. Allora

$$X = \frac{Z}{\sqrt{Y/n}} \quad (7.18)$$

ha distribuzione di Student con n gradi di libertà.

Si può richiedere ad R di eseguire direttamente il calcolo della densità di Student utilizzando la funzione

```
dt(x, df)
```

Gli argomenti di tale funzione sono

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria di Student;
- df numero di gradi di libertà (non negativo, può anche essere un valore non intero e $df = \text{Inf}$ è consentito).

Per calcolare la funzione di distribuzione invece utilizziamo la funzione

```
pt(x, df, lower.tail = TRUE)
```

¹Student è lo pseudonimo con cui il matematico inglese W.S. Gosset pubblicava i suoi articoli.

Gli argomenti di tale funzione sono

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria di Student;
- df numero di gradi di libertà;
- `lower.tail` se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

Il codice seguente

```
> curve(dnorm(x,mean=0,sd=1),from=-4, to=4,ylim=c(0,0.4),xlab="x",
+ylab="f(x)",main="n=1,2,5")
> curve(dt(x,df=1),add=TRUE,lty=2)
> text(0,0.29,"n=1")
>
> curve(dt(x,df=2),add=TRUE,lty=3)
> text(0,0.33,"n=2")
> curve(dt(x,df=5),add=TRUE,lty=4)
> arrows(0.2,0.37,1.0,0.39,code=1,length = 0.10)
> text(1.4,0.39,"n=5")
```

permette di rappresentare in Figura 7.16 la densità di probabilità di una variabile aleatoria $X \sim \mathcal{T}(n)$ per $n = 1, 2, 5$ e la densità normale standard (curva con tratto continuo). Notiamo che la densità (7.17) è unimodale, simmetrica intorno

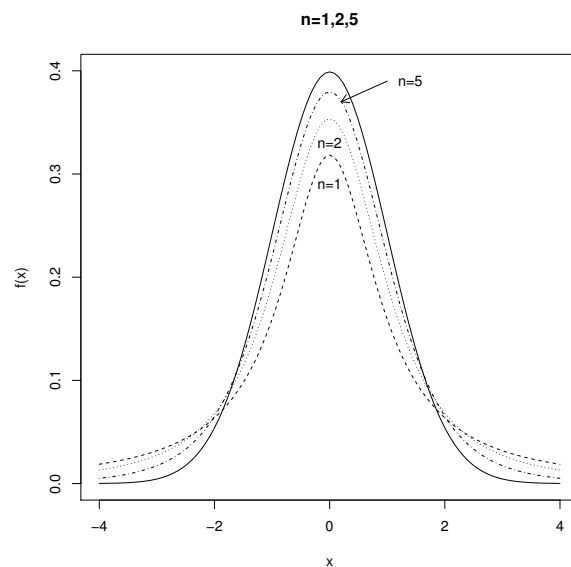


Figura 7.16: Densità di probabilità di $X \sim \mathcal{T}(n)$ per $n = 1, 2, 5$ e densità normale standard (curva con tratto continuo)

all'asse $x = 0$ e dipendente dal solo parametro rappresentante il numero di gradi

di libertà. Per $n = 1$ la (7.17) si identifica con la densità di Cauchy

$$f_X(x) = \frac{1}{\pi} \frac{1}{1+x^2}, \quad x \in \mathbb{R}$$

Per una variabile aleatoria di Cauchy il valore medio non esiste, mentre la moda e la mediana sono nulle.

Al limite per $n \rightarrow +\infty$ essa converge alla densità normale standard. Infatti, risulta

$$\lim_{n \rightarrow +\infty} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n} \pi \Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}, \quad x \in \mathbb{R}.$$

Il valore medio di $X \sim \mathcal{T}(n)$ non esiste se $n = 1$ (ossia quando la variabile aleatoria è di Cauchy) e risulta nullo se $n = 2, 3, \dots$. Inoltre la varianza non esiste se $n = 1$ e diverge se $n = 2$; per $n = 3, 4, \dots$ si ha poi:

$$\text{Var}(X) = \frac{n}{n-2}.$$

È anche possibile calcolare i quantili e simulare una variabile chi-quadrato tramite le funzioni

```
qt(p, df, lower.tail = TRUE)
rt(N, df)
```

dove `df` indica il numero di gradi di libertà; gli altri parametri sono già stati descritti per le altre distribuzioni.

La distribuzione di Student riveste un ruolo fondamentale nella stima puntuale ed intervallare del valore medio di una popolazione normale ed anche in molti test di verifica di ipotesi statistiche.

Nel seguito, utilizzeremo le distribuzioni di probabilità *normale*, *chi-quadrato* e di *Student* per la stima puntuale e intervallare dei parametri e nei test statistici di verifica di ipotesi sia per le distribuzioni di probabilità discrete che per le distribuzioni di probabilità continue.

Capitolo 8

Stima puntuale

8.1 Campioni casuali e stimatori

Uno dei problemi centrali dell'inferenza statistica è il seguente: *si desidera studiare una popolazione descritta da una variabile aleatoria osservabile X la cui funzione di distribuzione ha una forma nota ma contiene un parametro $\vartheta \in \Theta$ non noto (o più parametri non noti).*

Il termine *osservabile* significa che si possono osservare i valori assunti dalla variabile aleatoria X (ad esempio, eseguendo un esperimento casuale) e quindi il parametro non noto è presente soltanto nella legge di probabilità (funzione di distribuzione, funzione di probabilità, densità di probabilità). Ovviamente se ϑ è noto la legge di probabilità è completamente specificata.

Per ottenere informazioni sul parametro non noto ϑ della popolazione, si può fare uso dell'inferenza statistica considerando un campione estratto dalla popolazione e effettuando su tale campione delle opportune misure. Affinché le conclusioni dell'inferenza statistica siano valide il campione deve essere scelto in modo tale da essere *rappresentativo della popolazione*. Molti metodi dell'inferenza statistica sono basati sull'ipotesi di *campioni casuali*.

Definizione 8.1 *Si consideri una popolazione descritta da una variabile aleatoria osservabile X caratterizzata da funzione di distribuzione $F_X(x)$. Siano X_1, X_2, \dots, X_n delle variabili aleatorie osservabili indipendenti e identicamente distribuite (iid) con la stessa legge di probabilità della popolazione (ossia che costituiscono delle osservazioni di X). Il vettore aleatorio X_1, X_2, \dots, X_n è detto campione casuale di ampiezza n e la sua funzione di distribuzione è:*

$$\begin{aligned} F_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n) &= P(X_1 < x_1, X_2 < x_2, \dots, X_n < x_n) \\ &= P(X_1 < x_1) P(X_2 < x_2) \cdots P(X_n < x_n) = \prod_{i=1}^n F_X(x_i). \end{aligned} \quad (8.1)$$

Dalla Definizione 8.1 si nota che il campione casuale può essere estratto da una popolazione illimitata oppure da una popolazione finita; si suppone che l'estrazione avvenga con rimpiazzamento (per garantire l'indipendenza delle variabili aleatorie che costituiscono il campione).

Nei metodi di indagine dell'inferenza statistica si considera un campione casuale X_1, X_2, \dots, X_n di ampiezza n estratto dalla popolazione e si cerca di ottenere informazioni sul parametro non noto ϑ facendo uso di alcune variabili aleatorie, che sono funzioni misurabili del campione casuale, dette *statistiche* e *stimatori*.

Una statistica $t(X_1, X_2, \dots, X_n)$ è una funzione misurabile e osservabile del campione casuale X_1, X_2, \dots, X_n . Essendo la statistica osservabile, i valori da essa assunti dipendono soltanto dal campione osservato (x_1, x_2, \dots, x_n) estratto dalla popolazione e i parametri non noti sono presenti soltanto nella funzione di distribuzione della statistica.

Definizione 8.2 *Uno stimatore $\hat{\Theta} = t(X_1, X_2, \dots, X_n)$ è una funzione misurabile e osservabile del campione casuale X_1, X_2, \dots, X_n i cui valori possono essere usati per stimare un parametro non noto ϑ della popolazione. I valori $\hat{\vartheta}$ assunti da tale stimatore sono detti stime del parametro non noto ϑ .*

Statistiche tipiche sono la *media campionaria* e la *varianza campionaria*.

Definizione 8.3 *Sia X_1, X_2, \dots, X_n un campione casuale. La statistica*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (8.2)$$

è detta media campionaria, mentre la statistica

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (8.3)$$

è detta varianza campionaria.

Proposizione 8.1 *Sia X_1, X_2, \dots, X_n un campione casuale estratto da una popolazione descritta da una variabile aleatoria osservabile X caratterizzata da valore medio $E(X) = \mu$ finito e varianza $\text{Var}(X) = \sigma^2$ finita. Risulta:*

$$E(\bar{X}) = \mu, \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}. \quad (8.4)$$

Dimostrazione Per la proprietà di linearità del valore medio e l'identica distribuzione delle variabili aleatorie che costituiscono il campione, dalla (8.2) si ha:

$$E(\bar{X}) = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu.$$

Inoltre, poiché le variabili aleatorie che costituiscono il campione sono indipendenti ed identicamente distribuite, dalla (8.2) si ottiene:

$$\text{Var}(\bar{X}) = \text{Var}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{\sigma^2}{n}.$$

□

La Proposizione 8.1 mostra che al crescere dell'ampiezza del campione la media campionaria fornisce una stima sempre più accurata del valore medio della popolazione. Inoltre, dal teorema centrale di convergenza della probabilità scaturisce che per n sufficientemente grande (ossia per campioni di grande ampiezza) la funzione di distribuzione della media campionaria \bar{X} è approssimativamente normale con valore medio μ e varianza σ^2/n .

Proposizione 8.2 *Sia X_1, X_2, \dots, X_n un campione casuale estratto da una popolazione descritta da una variabile aleatoria osservabile X caratterizzata da valore medio $E(X) = \mu$, varianza $\text{Var}(X) = \sigma^2$ e avente i primi quattro momenti finiti. Risulta:*

$$E(S^2) = \sigma^2, \quad \text{Var}(S^2) = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \sigma^4 \right), \quad (8.5)$$

dove $\mu_4 = E(X^4)$.

Dimostrazione Dimostriamo per semplicità soltanto la prima delle (8.5). Osserviamo in primo luogo che

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X})^2 &= \sum_{i=1}^n [X_i - \mu - (\bar{X} - \mu)]^2 \\ &= \sum_{i=1}^n \left\{ (X_i - \mu)^2 + (\bar{X} - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) \right\} \\ &= \sum_{i=1}^n (X_i - \mu)^2 + n(\bar{X} - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu) \\ &= \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \end{aligned}$$

Ricordando la proprietà di linearità del valore medio e la definizione di varianza di una variabile aleatoria si ha:

$$\begin{aligned} E(S^2) &= E\left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right] = \frac{1}{n-1} E\left[\sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2\right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n \text{Var}(X_i) - n \text{Var}(\bar{X}) \right] = \frac{1}{n-1} \left[n\sigma^2 - n \frac{\sigma^2}{n} \right] = \sigma^2. \end{aligned}$$

□

8.2 Metodi per la ricerca di stimatori

Sia X_1, X_2, \dots, X_n un campione casuale estratto da una popolazione con funzione di probabilità (nel caso discreto) oppure densità di probabilità (nel caso assolutamente continuo) $f(x; \vartheta_1, \vartheta_2, \dots, \vartheta_k)$ dove $\vartheta_1, \vartheta_2, \dots, \vartheta_k$ denotano i parametri non noti della popolazione. Lo scopo del decisore, dopo aver osservato i valori assunti dal campione casuale, è quello di stimare i parametri non noti della popolazione. I principali metodi di stima puntuale dei parametri sono il *metodo dei momenti* e il *metodo della massima verosimiglianza*.

8.2.1 Metodo dei momenti

Il metodo dei momenti è uno dei più antichi metodi di stima dei parametri. Per illustrarlo occorre in primo luogo definire i *momenti campionari*.

Definizione 8.4 Si definisce momento campionario r -esimo relativo ai valori osservati (x_1, x_2, \dots, x_n) del campione casuale il valore

$$M_r(x_1, x_2, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i^r \quad (r = 1, 2, \dots) \quad (8.6)$$

Si nota quindi che il momento campionario r -esimo è la media aritmetica delle potenze r -esime delle n osservazioni effettuate sulla popolazione. In particolare, se $r = 1$ il momento campionario $M_1(x_1, x_2, \dots, x_n)$ coincide con il valore osservato della media campionaria \bar{X} , ossia $M_1 = (x_1 + x_2 + \dots + x_n)/n$.

Se esistono k parametri da stimare, il metodo dei momenti consiste nell'uguagliare i primi k momenti della popolazione in esame con i corrispondenti momenti del campione casuale. Quindi, se i primi k momenti esistono e sono finiti, tale metodo consiste nel risolvere il sistema di k equazioni

$$E(X^r) = M_r(x_1, x_2, \dots, x_n) \quad (r = 1, 2, \dots, k). \quad (8.7)$$

Le incognite del sistema sono i parametri $\vartheta_1, \vartheta_2, \dots, \vartheta_k$. Affinché il metodo dei momenti sia utilizzabile occorre che il sistema (8.7) ammetta un'unica soluzione. Le stime dei parametri ottenute con tale metodo, indicate con $\hat{\vartheta}_1, \hat{\vartheta}_2, \dots, \hat{\vartheta}_k$, dipendono dal campione osservato (x_1, x_2, \dots, x_n) e quindi al variare dei possibili campioni osservati si ottengono gli stimatori $\hat{\Theta}_1, \hat{\Theta}_2, \dots, \hat{\Theta}_k$ dei parametri non noti $\vartheta_1, \vartheta_2, \dots, \vartheta_k$ della popolazione, detti *stimatori del metodo dei momenti*. Alcune volte per ottenere tali stimatori è necessario utilizzare un numero maggiore di equazioni rispetto al numero dei parametri non noti da stimare.

► **(Popolazione di Bernoulli)** Ci si propone di determinare con il metodo dei momenti lo stimatore del parametro p di una popolazione di Bernoulli caratterizzata da funzione di probabilità

$$p_X(x) = p^x (1-p)^{1-x} \quad (x = 0, 1).$$

Occorre quindi stimare il parametro p . Poiché $E(X) = p$, dalla (8.7) si ha:

$$\hat{p} = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

Il metodo dei momenti fornisce quindi come stimatore del parametro p la media campionaria \bar{X} .

Se ad esempio, consideriamo un campione **campbern** di ampiezza 30 contenente i risultati di lanci indipendenti di una moneta

```
> campbern<-c(0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0,
+ 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1)
>
> stimap<-mean(campbern)
> stimap
[1] 0.5666667
```

la stima del parametro p con il metodo dei momenti è $\hat{p} = 0.567$. \diamond

► **(Popolazione geometrica)** Ci si propone di determinare con il metodo dei momenti lo stimatore del parametro p di una popolazione geometrica caratterizzata da funzione di probabilità

$$p_X(x) = \begin{cases} p(1-p)^{x-1}, & x = 1, 2, \dots \\ 0, & \text{altrimenti.} \end{cases}$$

Poiché $E(X) = 1/p$, ponendo $\vartheta = 1/p$ dalla (8.7) si ha:

$$\hat{\vartheta} = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

Il metodo dei momenti fornisce quindi come stimatore del parametro $\vartheta = 1/p$ la media campionaria \bar{X} .

Se ad esempio, consideriamo un campione **campgeom** di ampiezza 30 contenente i tempi di attesa per ottenere il primo successo in lanci ripetuti di una moneta

```
> campgeom<-c(5, 1, 6, 6, 2, 6, 3, 1, 1, 2, 8, 2, 5, 4, 4,
+ 2, 1, 6, 7, 1, 6, 4, 4, 3, 2, 3, 5, 2, 4, 3)
>
> stimap<-1/mean(campgeom)
> stimap
[1] 0.2752294
```

la stima del parametro p con il metodo dei momenti è $\hat{p} = 0.275$. \diamond

► **(Popolazione di Poisson)** Si desidera determinare con il metodo dei momenti lo stimatore del valore medio λ di una popolazione di Poisson avente funzione di probabilità:

$$p_X(x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda}, & x = 0, 1, \dots \quad (\lambda > 0), \\ 0, & \text{altrimenti.} \end{cases}$$

Occorre quindi stimare il parametro λ . Poiché $E(X) = \lambda$, dalla (8.7) si ha:

$$\hat{\lambda} = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

Il metodo dei momenti fornisce quindi come stimatore del parametro λ la media campionaria \overline{X} .

Se ad esempio, consideriamo un campione **camppois** di ampiezza 30 contenente il numero di utenti che arrivano ad un centro di calcolo in intervalli di 10 minuti

```
> camppois<-c(2, 2, 1, 2, 2, 1, 3, 3, 4, 1, 3, 6, 2, 2, 2,
+ 1, 2, 4, 2, 2, 6, 0, 1, 8, 2, 3, 4, 1, 1, 2)
>
> stimalambda<-mean(camppois)
> stimalambda
[1] 2.5
```

la stima del parametro λ con il metodo dei momenti è $\hat{\lambda} = 2.5$. \diamond

► **(Popolazione uniforme)** Si desidera determinare con il metodo dei momenti lo stimatore del parametro ϑ di una popolazione uniforme caratterizzata da funzione densità di probabilità

$$f_X(x) = \begin{cases} \frac{1}{\vartheta}, & 0 < x < \vartheta \\ 0, & \text{altrimenti.} \end{cases}$$

Occorre quindi stimare il parametro ϑ . Poiché $E(X) = \vartheta/2$, dalla (8.7) si ha:

$$\frac{\hat{\vartheta}}{2} = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

Il metodo dei momenti fornisce quindi come stimatore del parametro ϑ la media campionaria $2\overline{X}$.

Se ad esempio, riconsideriamo un campione **campunif** di ampiezza 30 contenente i tempi, misurati in minuti e supposti uniformi in un intervallo $(0, \vartheta)$, necessari per soddisfare le richieste di utenti che arrivano ad un centro di calcolo

```
> campunif<-c(1.556, 1.357, 1.574, 0.133, 1.748, 0.348, 0.566,
+ 0.767, 0.374, 1.856, 0.488, 0.327, 0.813, 0.005, 0.191, 1.311,
+ 0.345, 0.934, 0.140, 0.796, 0.254, 0.962, 1.318, 1.71, 0.257,
+ 0.605, 0.516, 0.083, 0.052, 0.290)
>
> stimatheta<-2.0*mean(campunif)
> stimatheta
[1] 1.445067
```

la stima del parametro ϑ con il metodo dei momenti è $\hat{\vartheta} = 1.445$. \diamond

► **(Popolazione esponenziale)** Si desidera determinare con il metodo dei momenti lo stimatore del valore medio $1/\lambda$ di una popolazione esponenziale avente densità di probabilità

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{altrimenti.} \end{cases}$$

Occorre quindi stimare il parametro $\vartheta = 1/\lambda$. Poiché $E(X) = 1/\lambda$, dalla (8.7) segue

$$\widehat{\vartheta} = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

Il metodo dei momenti fornisce quindi come stimatore del parametro $\vartheta = 1/\lambda$ la media campionaria \overline{X} .

Se ad esempio, consideriamo un campione `campexp` di ampiezza 30 contenente i tempi, supposti esponenziali, tra arrivi successivi (tempi di interarrivo) di utenti che arrivano ad un centro di calcolo

```
> campexp<-c(0.196, 0.409, 0.225, 0.224, 0.248, 0.280, 0.791,
+ 1.165, 0.355, 1.055, 0.393, 0.711, 0.455, 0.066, 0.179,
+ 0.543, 0.067, 0.635, 0.540, 1.454, 0.213, 0.532, 0.613, 1.876,
+ 0.047, 2.042, 0.018, 1.105, 0.098, 0.032)
>
> stimatheta<-1.0/mean(campexp)
> stimatheta
[1] 1.810829
```

la stima del parametro ϑ con il metodo dei momenti è $\widehat{\vartheta} = 1.81$. \diamond

► **(Popolazione normale)** Si è interessati a determinare con il metodo dei momenti gli stimatori dei parametri μ e σ^2 di una popolazione normale di densità di probabilità

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \quad (x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma > 0).$$

Occorre quindi stimare due parametri μ e σ^2 . Poiché $E(X) = \mu$ e $E(X^2) = \sigma^2 + \mu^2$, dalla (8.7) si ha:

$$\widehat{\mu} = \frac{x_1 + x_2 + \dots + x_n}{n}, \quad \widehat{\sigma^2} + \widehat{\mu^2} = \frac{x_1^2 + x_2^2 + \dots + x_n^2}{n},$$

da cui si ricava:

$$\begin{aligned} \widehat{\sigma^2} &= \frac{x_1^2 + x_2^2 + \dots + x_n^2}{n} - \frac{(x_1 + x_2 + \dots + x_n)^2}{n^2} = \frac{1}{n} \left[\sum_{i=1}^n x_i^2 - \overline{x} \sum_{i=1}^n x_i \right] \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - \overline{x})^2. \end{aligned}$$

Il metodo dei momenti fornisce quindi come stimatore del valore medio μ la media campionaria \overline{X} e come stimatore della varianza σ^2 la variabile aleatoria $(n-1)S^2/n$.

Se ad esempio, consideriamo un campione `campnorm` di ampiezza 30 contenente le lunghezze in metri, supposte normali, riscontrate nel misurare dei tubi prodotti da un'industria

```
> campnorm<-c(2.86, 3.03, 3.05, 3.32, 3.06, 2.91, 3.11, 3.21,
+ 2.85, 2.86, 2.78, 3.28, 3.39, 3.16, 3.05, 3.01, 3.10,
```

```

+ 2.88, 3.25, 2.89, 2.75, 2.99, 3.34, 2.93, 3.14, 2.99,
+ 2.97, 3.21, 3.27, 2.91)
>
> stimamu <- mean(campnorm)
> stimamu
[1] 3.052461
>
> stimasigma2 <- (length(campnorm)-1)*var(campnorm)/length(campnorm)
> stimasigma2
[1] 0.02996195

```

la stima del parametro μ con il metodo dei momenti è $\hat{\mu} = 3.052$ e la stima del parametro σ^2 con il metodo dei momenti è $\hat{\sigma}^2 = 0.03$. \diamond

8.2.2 Metodo della massima verosimiglianza

Il metodo della massima verosimiglianza è il più importante metodo per la stima dei parametri non noti di una popolazione e solitamente è preferito al metodo dei momenti. Per illustrare il metodo della massima verosimiglianza occorre introdurre in primo luogo la *funzione di verosimiglianza*.

Definizione 8.5 Sia X_1, X_2, \dots, X_n un campione casuale di ampiezza n estratto dalla popolazione. La funzione di verosimiglianza $L(\vartheta_1, \vartheta_2, \dots, \vartheta_k) = L(\vartheta_1, \vartheta_2, \dots, \vartheta_k; x_1, x_2, \dots, x_n)$ del campione osservato (x_1, x_2, \dots, x_n) è la funzione di probabilità congiunta (nel caso di popolazione discreta) oppure la funzione densità di probabilità congiunta (nel caso di popolazione assolutamente continua) del campione casuale X_1, X_2, \dots, X_n , ossia

$$\begin{aligned}
 L(\vartheta_1, \vartheta_2, \dots, \vartheta_k) &= L(\vartheta_1, \vartheta_2, \dots, \vartheta_k; x_1, x_2, \dots, x_n) \\
 &= f(x_1; \vartheta_1, \vartheta_2, \dots, \vartheta_k) f(x_2; \vartheta_1, \vartheta_2, \dots, \vartheta_k) \cdots f(x_n; \vartheta_1, \vartheta_2, \dots, \vartheta_k). \quad (8.8)
 \end{aligned}$$

Il metodo della massima verosimiglianza consiste nel massimizzare la funzione di verosimiglianza rispetto ai parametri $\vartheta_1, \vartheta_2, \dots, \vartheta_k$. Tale metodo cerca quindi di determinare da quale funzione di probabilità congiunta (nel caso di popolazione discreta) oppure di densità di probabilità congiunta (nel caso di popolazione assolutamente continua) è *più verosimile* (è *più plausibile*) che provenga il campione osservato (x_1, x_2, \dots, x_n) . Pertanto si cercano di determinare i valori $\vartheta_1, \vartheta_2, \dots, \vartheta_k$ che rendono massima la funzione di verosimiglianza e che quindi offrano, in un certo senso, la migliore spiegazione del campione osservato (x_1, x_2, \dots, x_n) .

I valori di $\vartheta_1, \vartheta_2, \dots, \vartheta_k$ che massimizzano la funzione di verosimiglianza sono indicati con $\hat{\vartheta}_1, \hat{\vartheta}_2, \dots, \hat{\vartheta}_k$; essi costituiscono le *stime di massima verosimiglianza* dei parametri non noti $\vartheta_1, \vartheta_2, \dots, \vartheta_k$ della popolazione. Tali stime dipendono dal campione osservato (x_1, x_2, \dots, x_n) e quindi al variare dei possibili campioni osservati si ottengono gli stimatori di massima verosimiglianza $\hat{\Theta}_1, \hat{\Theta}_2, \dots, \hat{\Theta}_k$ dei parametri non noti $\vartheta_1, \vartheta_2, \dots, \vartheta_k$ della popolazione, detti *stimatori di massima verosimiglianza*.

► **(Popolazione di Bernoulli)** Ci si propone di determinare lo stimatore di massima verosimiglianza del parametro p di una popolazione di Bernoulli

caratterizzata da funzione di probabilità

$$p_X(x) = p^x (1-p)^{1-x} \quad (x = 0, 1).$$

Si ha

$$\begin{aligned} L(p) &= p^{x_1} (1-p)^{1-x_1} p^{x_2} (1-p)^{1-x_2} \dots p^{x_n} (1-p)^{1-x_n} \\ &= p^{x_1+x_2+\dots+x_n} (1-p)^{n-(x_1+x_2+\dots+x_n)} \quad (0 < p < 1), \end{aligned}$$

dove le x_i possono assumere il valore 0 oppure il valore 1. Si nota che

$$\ln L(p) = \ln p \sum_{i=1}^n x_i + \left[n - \sum_{i=1}^n x_i \right] \ln(1-p) \quad (0 < p < 1)$$

da cui si ottiene:

$$\begin{aligned} \frac{d \ln L(p)}{dp} &= \frac{1}{p} \sum_{i=1}^n x_i - \frac{1}{1-p} \left[n - \sum_{i=1}^n x_i \right] = \frac{1}{p(1-p)} \sum_{i=1}^n x_i - \frac{n}{1-p} \\ &= \frac{n}{p(1-p)} \left[\frac{1}{n} \sum_{i=1}^n x_i - p \right]. \end{aligned}$$

La stima di massima verosimiglianza del parametro p è

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$$

e quindi lo stimatore di massima verosimiglianza di p è la media campionaria \bar{X} . \diamond

► **(Popolazione geometrica)** Si è interessati a determinare lo stimatore di massima verosimiglianza del parametro $1/p$ di una popolazione geometrica caratterizzata da funzione di probabilità

$$p_X(x) = \begin{cases} p(1-p)^{x-1}, & x = 1, 2, \dots \\ 0, & \text{altrimenti.} \end{cases}$$

Ponendo $\vartheta = 1/p$, si ha

$$\begin{aligned} L(\vartheta) &= \left(\frac{1}{\vartheta} \right)^n \left(1 - \frac{1}{\vartheta} \right)^{x_1+x_2+\dots+x_n-n} \\ &= \left(\frac{1}{\vartheta} \right)^{x_1+x_2+\dots+x_n} (\vartheta-1)^{x_1+x_2+\dots+x_n-n} \quad (\vartheta > 1) \end{aligned}$$

dove le x_i sono numeri interi positivi. Si nota che

$$\ln L(\vartheta) = -\ln \vartheta \sum_{i=1}^n x_i + \ln(\vartheta-1) \left(\sum_{i=1}^n x_i - n \right) \quad (\vartheta > 1)$$

e quindi si ricava

$$\begin{aligned}\frac{d \ln L(\vartheta)}{d \vartheta} &= -\frac{1}{\vartheta} \sum_{i=1}^n x_i + \frac{1}{\vartheta-1} \left(\sum_{i=1}^n x_i - n \right) \\ &= \left(-\frac{1}{\vartheta} + \frac{1}{\vartheta-1} \right) \sum_{i=1}^n x_i - \frac{n}{\vartheta-1} \\ &= \frac{1}{\vartheta(\vartheta-1)} \sum_{i=1}^n x_i - \frac{n}{\vartheta-1} = \frac{n}{\vartheta(\vartheta-1)} \left(\frac{1}{n} \sum_{i=1}^n x_i - \vartheta \right).\end{aligned}$$

La stima di massima verosimiglianza del parametro $\vartheta = 1/p$ è

$$\hat{\vartheta} = \frac{1}{n} \sum_{i=1}^n x_i$$

e quindi lo stimatore di massima verosimiglianza di $\vartheta = 1/p$ è la media campionaria \bar{X} . \diamond

► **(Popolazione di Poisson)** Si desidera determinare lo stimatore di massima verosimiglianza del parametro λ di una popolazione di Poisson caratterizzata da funzione di probabilità

$$P(X = x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \quad (x = 0, 1, \dots).$$

Si ha

$$L(\lambda) = \frac{\lambda^{x_1+x_2+\dots+x_n}}{x_1! x_2! \dots x_n!} e^{-n\lambda} \quad (\lambda > 0)$$

dove le x_i sono numeri interi non negativi. Si nota che

$$\ln L(\lambda) = \ln \lambda \sum_{i=1}^n x_i - n\lambda - \ln [x_1! x_2! \dots x_n!] \quad (\lambda > 0)$$

da cui segue

$$\frac{d \ln L(\lambda)}{d \lambda} = \frac{1}{\lambda} \sum_{i=1}^n x_i - n = \frac{n}{\lambda} \left(\frac{1}{n} \sum_{i=1}^n x_i - \lambda \right).$$

La stima di massima verosimiglianza del parametro λ è

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n x_i$$

e quindi lo stimatore di massima verosimiglianza di λ è la media campionaria \bar{X} . \diamond

► **(Popolazione uniforme)** Si desidera determinare lo stimatore di massima verosimiglianza del parametro ϑ di una popolazione uniforme caratterizzata da funzione densità di probabilità

$$f_X(x) = \begin{cases} \frac{1}{\vartheta}, & 0 < x < \vartheta \\ 0, & \text{altrimenti.} \end{cases}$$

Si ha

$$L(\vartheta) = \frac{1}{\vartheta^n} \quad (\vartheta > 0),$$

dove $0 < x_1 < \vartheta, 0 < x_2 < \vartheta, \dots, 0 < x_n < \vartheta$. Si nota che ϑ non può essere inferiore a nessuno dei dati del campione osservato, ossia $\vartheta > \max(x_1, x_2, \dots, x_n)$. Inoltre quando $\vartheta > \max(x_1, x_2, \dots, x_n)$, la funzione di verosimiglianza è strettamente monotona decrescente e quindi si può assumere la stima di massima verosimiglianza di ϑ è

$$\hat{\vartheta} = \max(x_1, x_2, \dots, x_n).$$

Lo stimatore di massima verosimiglianza del parametro ϑ è quindi $\hat{\Theta} = \max(X_1, X_2, \dots, X_n)$. Si noti che essendo $E(X) = \vartheta/2$, lo stimatore di ϑ ottenuto con il metodo dei momenti è invece $\hat{\Theta} = 2\bar{X}$ e differisce da quello ottenuto con il metodo della massima verosimiglianza.

Se ad esempio, riconsideriamo un campione *campunif* di ampiezza 30 contenente i tempi, misurati in ore e supposti uniformi in un intervallo $(0, \vartheta)$, necessari per soddisfare le richieste di utenti che arrivano ad un centro di calcolo

```
> campunif<-c(1.556, 1.357, 1.574, 0.133, 1.748, 0.348, 0.566,
+ 0.767, 0.374, 1.856, 0.488, 0.327, 0.813, 0.005, 0.191, 1.311,
+ 0.345, 0.934, 0.140, 0.796, 0.254, 0.962, 1.318, 1.71, 0.257,
+ 0.605, 0.516, 0.083, 0.052, 0.290)
>
> stimatheta<-max(campunif)
> stimatheta
[1] 1.856
```

la stima del parametro ϑ con il metodo della massima verosimiglianza è $\hat{\vartheta} = 1.856$. ◇

► **(Popolazione esponenziale)** Ci si propone di determinare lo stimatore di massima verosimiglianza del parametro $1/\lambda$ di una popolazione esponenziale con funzione densità di probabilità

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{altrimenti.} \end{cases}$$

Ponendo $\vartheta = 1/\lambda$, si ha

$$L(\vartheta) = \left(\frac{1}{\vartheta}\right)^n \exp\left\{-\frac{1}{\vartheta} \sum_{i=1}^n x_i\right\} \quad (\vartheta > 0)$$

dove le x_i sono positive. Si nota che

$$\ln L(\vartheta) = -n \ln \vartheta - \frac{1}{\vartheta} \sum_{i=1}^n x_i \quad (\vartheta > 0)$$

e quindi si ottiene

$$\frac{d \ln L(\vartheta)}{d\vartheta} = -\frac{n}{\vartheta} + \frac{1}{\vartheta^2} \sum_{i=1}^n x_i = \frac{n}{\vartheta^2} \left(\frac{1}{n} \sum_{i=1}^n x_i - \vartheta \right).$$

La stima di massima verosimiglianza del parametro $\vartheta = 1/\lambda$ è quindi:

$$\hat{\vartheta} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Lo stimatore di massima verosimiglianza di $1/\lambda$ è la media campionaria \overline{X} . \diamond

► **(Popolazione normale)** Si desidera determinare lo stimatore di massima verosimiglianza dei parametri μ e σ^2 di una popolazione normale caratterizzata da funzione densità di probabilità

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad (x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma > 0)$$

Si ha

$$L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp\left\{-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}\right\} \quad (\mu \in \mathbb{R}, \sigma > 0)$$

dove le $x_i \in \mathbb{R}$. Si nota che

$$\ln L(\mu, \sigma^2) = -\frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \quad (\mu \in \mathbb{R}, \sigma > 0)$$

e quindi si ha:

$$\begin{aligned} \frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = \frac{n}{\sigma^2} \left(\frac{1}{n} \sum_{i=1}^n x_i - \mu \right) \\ \frac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = -\frac{n}{2\sigma^4} \left(\sigma^2 - \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \right). \end{aligned}$$

Le stime di massima verosimiglianza dei parametri μ e σ^2 sono rispettivamente

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2.$$

Lo stimatore di massima verosimiglianza di μ è la media campionaria \overline{X} . Invece lo stimatore di σ^2 è $(n-1)S^2/n$. Entrambi gli stimatori coincidono con quelli ottenuti con il metodo dei momenti.

\diamond

8.3 Proprietà degli stimatori

In generale esistono molti stimatori che possono essere utilizzati per stimare il parametro non noto di una popolazione. Occorre quindi definire delle proprietà di cui può o meno godere uno stimatore. Alcune di queste proprietà sono:

- *corretto* (o equivalentemente *non distorto*),
- *più efficiente di un altro*,
- *corretto e con varianza uniformemente minima*,
- *asintoticamente corretto*,
- *consistente*.

Definizione 8.6 Uno stimatore $\hat{\Theta} = t(X_1, X_2, \dots, X_n)$ del parametro non noto ϑ della popolazione è detto *corretto* (non distorto) se e solo se per ogni $\vartheta \in \Theta$ si ha

$$E(\hat{\Theta}) = \vartheta, \quad (8.9)$$

ossia se il valore medio dello stimatore $\hat{\Theta}$ è uguale al corrispondente parametro non noto della popolazione.

Occorre sottolineare che possono esistere differenti stimatori corretti di un parametro non noto di una popolazione.

Dalle Proposizioni 8.1 e 8.2 segue che se X_1, X_2, \dots, X_n è un campione casuale di ampiezza n estratto da una popolazione caratterizzata da valore medio μ e varianza σ^2 , allora la media campionaria \bar{X} e la varianza campionaria S^2 sono rispettivamente stimatori corretti del valore medio μ e della varianza σ^2 della popolazione.

► **(Popolazione di Bernoulli)** Sia X_1, X_2, \dots, X_n un campione casuale di ampiezza n estratto da una popolazione di Bernoulli caratterizzata da funzione di probabilità

$$p_X(x) = p^x (1-p)^{1-x} \quad (x = 0, 1).$$

Poiché $E(\bar{X}) = E(X) = p$, la media campionaria \bar{X} , individuata sia con il metodo dei momenti che con il metodo della massima verosimiglianza, è uno stimatore corretto del parametro non noto p della popolazione. \diamond

Esistono vari stimatori per uno stesso parametro non noto ϑ di una popolazione. Occorre quindi definire alcuni criteri che permettano di *confrontare più stimatori dello stesso parametro*.

La dispersione di uno stimatore rispetto al parametro non noto ϑ può essere misurata in vari modi. Una misura molto importante è *l'errore quadratico medio*, che fornisce una misura di quanto si discosta lo stimatore $\hat{\Theta}$ dal parametro non noto ϑ della popolazione.

Definizione 8.7 Sia $\hat{\Theta} = t(X_1, X_2, \dots, X_n)$ uno stimatore del parametro non noto ϑ della popolazione. Si chiama *errore quadratico medio* la quantità

$$MSE(\hat{\Theta}) = E[(\hat{\Theta} - \vartheta)^2]. \quad (8.10)$$

Il principale problema del decisore consiste nello scegliere lo stimatore migliore del parametro ϑ , ossia lo stimatore che ha il più piccolo errore quadratico medio per ogni valore ammissibile di $\vartheta \in \Theta$. Situazioni in cui esiste uno stimatore migliore di tutti gli altri si verificano raramente e spesso sono poco interessanti. La ricerca dello stimatore con errore quadratico uniformemente minimo deve essere quindi effettuata in opportune classi come, ad esempio, nella *classe degli stimatori corretti*.

Proposizione 8.3 *Se $\hat{\Theta}$ è uno stimatore corretto del parametro ϑ , allora*

$$MSE(\hat{\Theta}) = E\{[\hat{\Theta} - E(\hat{\Theta})]^2\} = \text{Var}(\hat{\Theta}) \quad (8.11)$$

La Proposizione 8.3 mostra che se si restringe la ricerca alla classe degli stimatori corretti del parametro non noto ϑ , il problema del decisore consiste nel determinare in tale classe uno *stimatore con varianza uniformemente minima*.

Definizione 8.8 *Uno stimatore $\hat{\Theta}$ si dice corretto con varianza uniformemente minima per il parametro non noto ϑ se e solo se per ogni $\vartheta \in \Theta$ risulta*

$$(i) \ E(\hat{\Theta}) = \vartheta,$$

$$(ii) \ \text{Var}(\hat{\Theta}) \leq \text{Var}(\hat{\Theta}^*) \text{ per ogni altro stimatore } \hat{\Theta}^* \text{ corretto del parametro } \vartheta.$$

La varianza fornisce quindi una misura della dispersione dei valori assunti dallo stimatore intorno al suo valore medio. Nella ricerca di uno stimatore corretto con varianza uniformemente minima è spesso utilizzata la seguente disuguaglianza.

Proposizione 8.4 (*Disuguaglianza di Cramér–Rao*) *Sia $\hat{\Theta} = t(X_1, X_2, \dots, X_n)$ uno stimatore corretto del parametro non noto ϑ di una popolazione caratterizzata da funzione di probabilità (nel caso discreto) oppure densità di probabilità (nel caso assolutamente continuo) $f(x; \vartheta)$. Se sono soddisfatte le seguenti ipotesi*

$$(a) \ \frac{\partial}{\partial \vartheta} \ln f(x; \vartheta) \text{ esiste per ogni } x \text{ e per ogni } \vartheta \in \Theta,$$

$$(b) \ E\left\{\left[\frac{\partial}{\partial \vartheta} \ln f(X; \vartheta)\right]^2\right\} \text{ esiste finito per ogni } \vartheta \in \Theta,$$

la varianza dello stimatore $\hat{\Theta}$ soddisfa la disuguaglianza

$$\text{Var}(\hat{\Theta}) \geq \frac{1}{nE\left\{\left[\frac{\partial}{\partial \vartheta} \ln f(X; \vartheta)\right]^2\right\}}. \quad (8.12)$$

Si noti che la disuguaglianza di Cramér–Rao (8.12) individua l'estremo inferiore della varianza di uno stimatore corretto, ma non implica che esista sempre uno stimatore con varianza uguale al suo estremo.

Proposizione 8.5 *Nelle ipotesi della Proposizione 8.4, se*

$$\text{Var}(\hat{\Theta}) = \frac{1}{nE\left\{\left[\frac{\partial}{\partial\vartheta}\ln f(X;\vartheta)\right]^2\right\}}, \quad (8.13)$$

allora $\hat{\Theta}$ è uno stimatore corretto con varianza uniformemente minima per il parametro ϑ .

► **(Popolazione di Poisson)** Si desidera verificare che \bar{X} è uno stimatore corretto con varianza uniformemente minima del valore medio di una popolazione di Poisson di parametro $\lambda > 0$. Tale stimatore è stato determinato sia con il metodo dei momenti che con il metodo della massima verosimiglianza.

La funzione di probabilità è:

$$p(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \quad (x = 0, 1, \dots).$$

Poiché $E(X) = \lambda$, il parametro da stimare è $\vartheta = \lambda$. Se $x = 0, 1, \dots$ si ha

$$\ln f(x; \lambda) = -\ln x! + x \ln \lambda - \lambda$$

e quindi

$$\frac{\partial}{\partial\lambda} \ln f(x; \lambda) = \frac{x}{\lambda} - 1 = \frac{x - \lambda}{\lambda}.$$

Essendo $\text{Var}(X) = \lambda$, si ottiene:

$$E\left\{\left[\frac{\partial}{\partial\lambda} \ln f(X; \lambda)\right]^2\right\} = E\left[\left(\frac{X - \lambda}{\lambda}\right)^2\right] = \frac{1}{\lambda^2} E[(X - \lambda)^2] = \frac{\text{Var}(X)}{\lambda^2} = \frac{1}{\lambda}$$

e quindi

$$\text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n} = \frac{\lambda}{n}, \quad \frac{1}{nE\left\{\left[\frac{\partial}{\partial\lambda} \ln f(X; \lambda)\right]^2\right\}} = \frac{\lambda}{n}.$$

Dalla Proposizione 8.5 risulta quindi che \bar{X} è uno stimatore corretto con varianza uniformemente minima del valore medio λ di una popolazione di Poisson. \diamond

► **(Popolazione esponenziale)** Si è interessati a verificare che \bar{X} è uno stimatore corretto con varianza uniformemente minima del valore medio $1/\lambda$ ($\lambda > 0$) di una popolazione esponenziale. Tale stimatore è stato determinato sia con il metodo dei momenti che con il metodo della massima verosimiglianza.

La funzione densità di probabilità esponenziale è:

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0. \end{cases}$$

Poiché $E(X) = 1/\lambda$, il parametro da stimare è $\vartheta = 1/\lambda$. Se $x > 0$ si ha

$$\ln f(x; \vartheta) = \ln \lambda - \lambda x = -\ln \vartheta - \frac{x}{\vartheta}$$

e quindi

$$\frac{\partial}{\partial \vartheta} \ln f(x; \vartheta) = -\frac{1}{\vartheta} + \frac{x}{\vartheta^2} = \frac{x - \vartheta}{\vartheta^2}.$$

Poiché $\text{Var}(X) = 1/\lambda^2 = \vartheta^2$, si ottiene:

$$E\left\{\left[\frac{\partial}{\partial \vartheta} \ln f(X; \vartheta)\right]^2\right\} = E\left[\left(\frac{X - \vartheta}{\vartheta^2}\right)^2\right] = \frac{1}{\vartheta^4} E[(X - \vartheta)^2] = \frac{\text{Var}(X)}{\vartheta^4} = \frac{1}{\vartheta^2}$$

e quindi

$$\text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n} = \frac{1}{n\lambda^2}, \quad \frac{1}{nE\left\{\left[\frac{\partial}{\partial \vartheta} \ln f(X; \vartheta)\right]^2\right\}} = \frac{\vartheta^2}{n} = \frac{1}{n\lambda^2}.$$

Dalla Proposizione 8.5 segue che \bar{X} è uno stimatore corretto con varianza uniformemente minima del valore medio $1/\lambda$ di una popolazione esponenziale. \diamond

► **(Popolazione normale)** Si desidera verificare che \bar{X} è uno stimatore corretto con varianza uniformemente minima del valore medio μ di una popolazione normale avente varianza nota σ^2 . Tale stimatore è stato precedentemente determinato sia con il metodo dei momenti che con il metodo della massima verosimiglianza.

La densità di probabilità che caratterizza la popolazione è:

$$f(x; \mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\} \quad (x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma > 0).$$

Poiché $E(X) = \mu$, il parametro da stimare è $\vartheta = \mu$. Osserviamo che

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

e quindi

$$\frac{\partial}{\partial \mu} \ln f(x; \mu) = \frac{x - \mu}{\sigma^2}.$$

Essendo $\text{Var}(X) = \sigma^2$ risulta:

$$E\left\{\left[\frac{\partial}{\partial \mu} \ln f(X; \mu)\right]^2\right\} = E\left[\left(\frac{X - \mu}{\sigma^2}\right)^2\right] = \frac{1}{\sigma^4} E[(X - \mu)^2] = \frac{\text{Var}(X)}{\sigma^4} = \frac{1}{\sigma^2}$$

e quindi

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}, \quad \frac{1}{nE\left\{\left[\frac{\partial}{\partial \mu} \ln f(X; \mu)\right]^2\right\}} = \frac{\sigma^2}{n}.$$

Dalla Proposizione 8.5 segue quindi che \bar{X} è uno stimatore corretto con varianza uniformemente minima del valore medio μ di una popolazione normale con varianza nota σ^2 . \diamond

Occorre sottolineare che la media campionaria \bar{X} non è sempre uno stimatore corretto con varianza uniformemente minima del valore medio di una qualsiasi popolazione.

Abbiamo finora supposto che il campione casuale avesse una ampiezza fissa. Consideriamo ora una successione di campioni casuali (X_1) , (X_1, X_2) , \dots , (X_1, X_2, \dots, X_n) , \dots di ampiezze crescenti e definiamo una successione $\hat{\Theta}_1 = t(X_1)$, $\hat{\Theta}_2 = t(X_1, X_2)$, \dots , $\hat{\Theta}_n = t(X_1, X_2, \dots, X_n)$, \dots di stimatori del parametro non noto ϑ della popolazione. Spesso si desidera che al crescere dell'ampiezza del campione tali stimatori forniscano stime sempre più accurate del parametro ϑ . Per campioni di grande ampiezza alcune proprietà asintotica di uno stimatore sono la *correttezza asintotica* e la *consistenza*.

Definizione 8.9 *Uno stimatore $\hat{\Theta}_n = t(X_1, X_2, \dots, X_n)$ del parametro non noto ϑ della popolazione è detto asintoticamente corretto (asintoticamente non distorto) se e solo se per ogni $\vartheta \in \Theta$ si ha*

$$\lim_{n \rightarrow +\infty} E(\hat{\Theta}_n) = \vartheta, \quad (8.14)$$

ossia se il valore medio dello stimatore Θ_n tende al crescere dell'ampiezza del campione casuale al corrispondente parametro non noto della popolazione.

► Si desidera verificare che

$$\hat{\Theta}_n = \frac{n-1}{n} S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

è uno stimatore asintoticamente corretto del parametro σ^2 di una popolazione.

Ricordando che $E(S^2) = \sigma^2$, si ottiene immediatamente:

$$\lim_{n \rightarrow +\infty} E(\hat{\Theta}_n) = \lim_{n \rightarrow +\infty} \frac{n-1}{n} E(S^2) = \sigma^2 \lim_{n \rightarrow +\infty} \frac{n-1}{n} = \sigma^2.$$

In particolare, per una **popolazione normale** lo stimatore $(n-1)S^2/n$ della varianza σ^2 , individuato sia con il metodo dei momenti che con il metodo della massima verosimiglianza, è asintoticamente corretto. \diamond

► (**Popolazione di Bernoulli**) Per una popolazione di Bernoulli di parametro p si desidera verificare che

$$\hat{\Theta}_n = \frac{n\bar{X} + 1}{n + 2}$$

è uno stimatore asintoticamente corretto del parametro p .

Ricordando che $E(X) = p$, si ricava immediatamente che

$$\lim_{n \rightarrow +\infty} E(\hat{\Theta}_n) = \lim_{n \rightarrow +\infty} E\left(\frac{n\bar{X} + 1}{n + 2}\right) = \lim_{n \rightarrow +\infty} \frac{nE(\bar{X}) + 1}{n + 2} = \lim_{n \rightarrow +\infty} \frac{np + 1}{n + 2} = p.$$

\diamond

Vediamo ora il significato dell'altra proprietà asintotica, ossia della *consistenza*.

Definizione 8.10 Uno stimatore $\hat{\Theta}_n = t(X_1, X_2, \dots, X_n)$ del parametro non noto ϑ della popolazione è detto consistente se e solo se per ogni $\varepsilon > 0$ si ha

$$\lim_{n \rightarrow +\infty} P(|\hat{\Theta}_n - \vartheta| < \varepsilon) = 1 \quad \forall \vartheta \in \Theta,$$

ossia se e solo se $\hat{\Theta}_n$ converge in probabilità a ϑ .

Se la popolazione ha valore medio $E(X) = \mu$ finito, allora dalla legge debole dei grandi numeri di Khintchin della probabilità si ricava che la media campionaria \bar{X} è uno stimatore consistente per μ . Una condizione sufficiente affinché uno stimatore sia consistente è fornita nella seguente proposizione.

Proposizione 8.6 Lo stimatore $\hat{\Theta}_n = t(X_1, X_2, \dots, X_n)$ del parametro non noto ϑ della popolazione è consistente se

$$i) \lim_{n \rightarrow \infty} E(\hat{\Theta}_n) = \vartheta, \quad \forall \vartheta \in \Theta,$$

$$ii) \lim_{n \rightarrow +\infty} \text{Var}(\hat{\Theta}_n) = 0, \quad \forall \vartheta \in \Theta.$$

Si noti che la Proposizione 8.6 fornisce una condizione sufficiente ma non necessaria. Infatti uno stimatore può essere consistente senza essere asintoticamente corretto.

Uno stimatore può possedere o meno le proprietà precedentemente descritte e tali proprietà non sempre coesistono per uno stesso stimatore. La scelta dello stimatore, e quindi delle sue proprietà, deve essere effettuata da un decisore e dipende fondamentalmente dalla natura dell'indagine statistica.

Occorre infine sottolineare che sotto condizioni non molto restrittive il *metodo della massima verosimiglianza*, nel caso in cui esista un unico parametro ϑ da stimare, permette di determinare stimatori che godono di importanti proprietà asintotiche. Infatti lo stimatore del parametro ϑ che si ottiene con il metodo della massima verosimiglianza è asintoticamente corretto e consistente.

► **(Popolazione uniforme)** Consideriamo una popolazione uniforme caratterizzata da funzione densità di probabilità

$$f_X(x) = \begin{cases} \frac{1}{\vartheta}, & 0 < x < \vartheta \\ 0, & \text{altrimenti.} \end{cases}$$

Abbiamo precedentemente mostrato che lo stimatore di ϑ ottenuto con il metodo dei momenti è $\hat{\Theta} = 2\bar{X}$, mentre quello ottenuto con il metodo della massima verosimiglianza è $\hat{\Theta} = \max(X_1, X_2, \dots, X_n)$. Desideriamo analizzare le proprietà di questi due stimatori e definire un ulteriore stimatore corretto più efficiente.

- **Stimatore di ϑ con il metodo dei momenti.** Vogliamo mostrare che $\hat{\Theta} = 2\bar{X}$ è corretto e consistente. Infatti, si ha

$$E(\hat{\Theta}) = 2E(\bar{X}) = 2E(X) = 2 \frac{\vartheta}{2} = \vartheta,$$

$$\text{Var}(\hat{\Theta}) = 4\text{Var}(\bar{X}) = 4 \frac{\text{Var}(X)}{n} = \frac{4}{n} \frac{\vartheta^2}{12} = \frac{\vartheta^2}{3n},$$

da cui segue che lo stimatore $\hat{\Theta} = 2\bar{X}$ del parametro ϑ è corretto e consistente, essendo $\lim_{n \rightarrow +\infty} \text{Var}(\hat{\Theta}) = 0$.

- **Stimatore di ϑ con il metodo della massima verosimiglianza.** Vogliamo mostrare che $\hat{\Theta} = \max(X_1, X_2, \dots, X_n)$ è asintoticamente corretto e consistente. Osserviamo che per l'indipendenza e l'identica distribuzione delle variabili del campione casuale si ha:

$$\begin{aligned} P(\hat{\Theta} \leq x) &= P(\max(X_1, X_2, \dots, X_n) \leq x) = P(X_1 \leq x, X_2 \leq x, \dots, X_n \leq x) \\ &= P(X_1 \leq x) P(X_2 \leq x) \cdots P(X_n \leq x) = \begin{cases} 0, & x < 0 \\ \left(\frac{x}{\vartheta}\right)^n, & 0 \leq x < \vartheta \\ 1, & x \geq \vartheta. \end{cases} \end{aligned}$$

da cui la densità dello stimatore $\hat{\Theta} = \max(X_1, X_2, \dots, X_n)$ è

$$f(x) = \begin{cases} n\vartheta^{-n}x^{n-1}, & 0 < x < \vartheta \\ 0, & \text{altrimenti.} \end{cases}$$

Si ricava quindi che

$$\begin{aligned} E(\hat{\Theta}) &= \int_0^{\vartheta} x f(x) dx = \frac{n}{n+1} \vartheta, \\ E(\hat{\Theta}^2) &= \int_0^{\vartheta} x^2 f(x) dx = \frac{n}{n+2} \vartheta^2 \\ \text{Var}(\hat{\Theta}) &= E(\hat{\Theta}^2) - [E(\hat{\Theta})]^2 = \frac{n}{(n+2)(n+1)^2} \vartheta^2, \end{aligned}$$

da cui segue che lo stimatore $\hat{\Theta} = \max(X_1, X_2, \dots, X_n)$ del parametro ϑ è asintoticamente corretto e consistente, essendo $\lim_{n \rightarrow +\infty} \text{Var}(\hat{\Theta}) = 0$.

- **Altro stimatore corretto di ϑ .** Consideriamo infine un altro stimatore corretto di ϑ

$$\hat{\Theta}_1 = \frac{n+1}{n} \max(X_1, X_2, \dots, X_n)$$

Tale stimatore gode delle seguenti proprietà:

$$\begin{aligned} E(\hat{\Theta}_1) &= \frac{n+1}{n} E[\max(X_1, X_2, \dots, X_n)] = \vartheta, \\ \text{Var}(\hat{\Theta}_1) &= \frac{(n+1)^2}{n^2} \text{Var}[\max(X_1, X_2, \dots, X_n)] = \frac{1}{n(n+2)} \vartheta^2, \end{aligned}$$

ossia è corretto e consistente. Essendo la varianza di questo stimatore minore o uguale della varianza dello stimatore determinato con il metodo dei momenti, ne segue che lo stimatore $(n+1) \max(X_1, X_2, \dots, X_n)/n$ è più efficiente di $\hat{\Theta} = 2\bar{X}$. \diamond

Questo esempio mostra che nella stima di un parametro non noto di una popolazione si possono utilizzare diversi tipi di stimatori; la scelta spetta al decisore e si basa sull'ampiezza del campione e sulla natura dell'indagine statistica.

Nella Tabella 8.1 riassumiamo le proprietà di alcuni stimatori precedentemente considerati.

Tabella 8.1: Alcune proprietà degli stimatori

X	Metodo dei momenti	Metodo della massima verosimiglianza	Proprietà degli stimatori
Bernoulli $E(X) = p$	\bar{X}	\bar{X}	Stimatore corretto con varianza minima e consistente per p
Geometrica $E(X) = 1/p$	\bar{X}	\bar{X}	Stimatore corretto con varianza minima e consistente per $1/p$
Poisson $E(X) = \lambda$	\bar{X}	\bar{X}	Stimatore corretto con varianza minima e consistente per λ
Uniforme in $(0, \vartheta)$ $E(X) = \vartheta/2$	(1) \bar{X}	(2) $\max(X_1, X_2, \dots, X_n)/2$	(1) Stimatore corretto e consistente per $\vartheta/2$ (2) Stimatore asintoticamente corretto e consistente per $\vartheta/2$
Esponeziale $E(X) = 1/\lambda$	\bar{X}	\bar{X}	Stimatore corretto con varianza minima e consistente per $1/\lambda$
Normale $E(X) = \mu$ $Var(X) = \sigma^2$	(*) \bar{X} (**) $(n-1)S^2/n$	\bar{X} $(n-1)S^2/n$	(*) Stimatore corretto con varianza minima e consistente per μ (**) Stimatore asintoticamente corretto e consistente per σ^2

Capitolo 9

Intervalli di confidenza

9.1 Intervalli di confidenza

Alla stima puntuale di un parametro non noto di una popolazione (costituita da un singolo valore reale) spesso si preferisce sostituire un intervallo di valori, detto *intervallo di confidenza* (o intervallo di fiducia), ossia si cerca di determinare in base ai dati del campione, due limiti (uno inferiore ed uno superiore) entro i quali sia compreso il parametro non noto con un certo *coefficiente di confidenza* (detto anche *grado di fiducia*).

Sia X_1, X_2, \dots, X_n un campione casuale di ampiezza n estratto da una popolazione con funzione di probabilità (nel caso discreto) oppure densità di probabilità (nel caso assolutamente continuo) $f(x; \vartheta)$, dove ϑ denota il parametro non noto della popolazione. Denotiamo con $\underline{C}_n = g_1(X_1, X_2, \dots, X_n)$ e con $\overline{C}_n = g_2(X_1, X_2, \dots, X_n)$ due statistiche (funzioni osservabili del campione casuale) che soddisfino la condizione $\underline{C}_n < \overline{C}_n$, cioè che godono della proprietà che per ogni possibile fissato campione osservato $\mathbf{x} = (x_1, x_2, \dots, x_n)$ risulti $g_1(\mathbf{x}) < g_2(\mathbf{x})$.

Definizione 9.1 Fissato un coefficiente di confidenza $1 - \alpha$ ($0 < \alpha < 1$), se è possibile scegliere le statistiche \underline{C}_n e \overline{C}_n in modo tale che

$$P(\underline{C}_n < \vartheta < \overline{C}_n) = 1 - \alpha, \quad (1.1)$$

allora si dice che $(\underline{C}_n, \overline{C}_n)$ è un intervallo di confidenza (intervallo di fiducia) di grado $1 - \alpha$ per ϑ . Inoltre, le statistiche \underline{C}_n e \overline{C}_n sono dette limite inferiore e superiore dell'intervallo di confidenza.

Se $g_1(\mathbf{x})$ e $g_2(\mathbf{x})$ sono i valori assunti dalle statistiche \underline{C}_n e \overline{C}_n per il campione osservato $\mathbf{x} = (x_1, x_2, \dots, x_n)$, allora l'intervallo $(g_1(\mathbf{x}), g_2(\mathbf{x}))$ è detto *stima dell'intervallo di confidenza* di grado $1 - \alpha$ per ϑ ed i punti finali $g_1(\mathbf{x})$ e $g_2(\mathbf{x})$ di tale intervallo sono detti rispettivamente *stima del limite inferiore* e *stima del limite superiore dell'intervallo di confidenza*.

In generale esistono numerosi intervalli di confidenza dello stesso grado $1 - \alpha$ per un parametro non noto ϑ della popolazione. La scelta dell'intervallo di confidenza deve essere effettuata in base ad alcune proprietà statistiche. Ad esempio, fissato un coefficiente di confidenza $1 - \alpha$, alcune proprietà desiderabili sono che la *lunghezza dell'intervallo di confidenza*

$$L(X_1, X_2, \dots, X_n; 1 - \alpha) = \overline{C}_n - \underline{C}_n \quad (9.1)$$

sia la più piccola possibile oppure che la lunghezza media di tale intervallo sia la più piccola possibile.

Metodo pivotale

Un metodo per la costruzione degli intervalli di confidenza è il *metodo pivotale*. Tale metodo consiste essenzialmente nel determinare una variabile aleatoria di pivot $\gamma(X_1, X_2, \dots, X_n; \vartheta)$ che dipende dal campione casuale X_1, X_2, \dots, X_n e dal parametro non noto ϑ e la cui *funzione di distribuzione non contiene il parametro da stimare*. Tale variabile aleatoria non è una statistica poiché dipende dal parametro non noto ϑ e quindi non è osservabile.

Per ogni fissato coefficiente α ($0 < \alpha < 1$) siano α_1 e α_2 ($\alpha_1 < \alpha_2$) due valori dipendenti soltanto dal coefficiente fissato α tali che per ogni $\vartheta \in \Theta$ si abbia:

$$P(\alpha_1 < \gamma(X_1, X_2, \dots, X_n; \vartheta) < \alpha_2) = 1 - \alpha. \quad (9.2)$$

Se per ogni possibile campione osservato (x_1, x_2, \dots, x_n) e per ogni $\vartheta \in \Theta$, si riesce a dimostrare che

$$\alpha_1 < \gamma(\mathbf{x}; \vartheta) < \alpha_2 \iff g_1(\mathbf{x}) < \vartheta < g_2(\mathbf{x})$$

con $g_1(\mathbf{x})$ e $g_2(\mathbf{x})$ dipendenti soltanto dal campione osservato, allora la (9.2) è equivalente a richiedere che

$$P(g_1(X_1, X_2, \dots, X_n) < \vartheta < g_2(X_1, X_2, \dots, X_n)) = 1 - \alpha.$$

Denotando con $\underline{C}_n = g_1(X_1, X_2, \dots, X_n)$ e $\overline{C}_n = g_2(X_1, X_2, \dots, X_n)$, dalla Definizione 9.1 segue che $(\underline{C}_n, \overline{C}_n)$ è un intervallo di confidenza di grado $1 - \alpha$ per il parametro non noto ϑ della popolazione.

Occorre osservare che la (9.2) è senza dubbio soddisfatta se per ogni campione osservato $\mathbf{x} = (x_1, x_2, \dots, x_n)$ e per ogni $\vartheta \in \Theta$ risulta che $\gamma(\mathbf{x}; \vartheta)$ è una funzione strettamente monotona in ϑ .

9.2 Popolazione normale

Sia X_1, X_2, \dots, X_n un campione casuale di ampiezza n estratto da una popolazione normale con valore medio μ e varianza σ^2 . Si possono analizzare i seguenti problemi:

- (i) determinare un intervallo di confidenza di grado $1 - \alpha$ per il valore medio μ nel caso in cui la varianza σ^2 della popolazione normale è nota;
- (ii) determinare un intervallo di confidenza di grado $1 - \alpha$ per il valore medio μ nel caso in cui la varianza della popolazione normale è non nota;
- (iii) determinare un intervallo di confidenza di grado $1 - \alpha$ per la varianza σ^2 nel caso in cui il valore medio μ della popolazione normale è noto;
- (iv) determinare un intervallo di confidenza di grado $1 - \alpha$ per la varianza σ^2 nel caso in cui il valore medio della popolazione normale è non noto.

► **(Intervallo di confidenza per μ con σ^2 nota)**

Sia

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$$

la media campionaria. Abbiamo mostrato nel Cap. 8 che tale statistica gode delle seguenti proprietà

$$E(\bar{X}_n) = \mu, \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Per determinare un intervallo di confidenza di grado $1 - \alpha$ per il valore medio μ nel caso in cui la varianza σ^2 della popolazione normale è nota, utilizziamo il metodo pivotale e consideriamo la variabile aleatoria

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}},$$

che è una variabile aleatoria standardizzata (di valor medio nullo e varianza unitaria). Tale variabile aleatoria dipende dal campione casuale e dal parametro non noto μ (la varianza σ^2 è nota) e, quindi, può essere interpretata come una variabile aleatoria di pivot. Inoltre, essendo

$$Z_n = \frac{X_1 + X_2 + \dots + X_n - n\mu}{\sigma\sqrt{n}},$$

tale variabile aleatoria è distribuita normalmente con valore medio nullo e varianza unitaria, ossia è una normale standard. Scegliendo nel metodo pivotale $\alpha_1 = -z_{\alpha/2}$ e $\alpha_2 = z_{\alpha/2}$, dove $z_{\alpha/2}$ è tale che

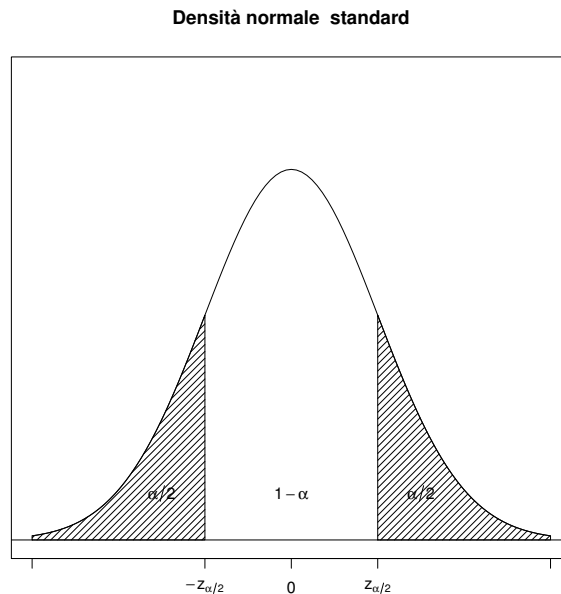
$$P(Z_n < -z_{\alpha/2}) = P(Z_n > z_{\alpha/2}) = \frac{\alpha}{2},$$

si ha:

$$P(-z_{\alpha/2} < Z_n < z_{\alpha/2}) = 1 - \alpha. \quad (9.3)$$

Ciò è evidenziato in Figura 9.1 ottenuta tramite il seguente codice:

A.G. Nobile

Figura 9.1: Densità normale standard e grado di fiducia $1 - \alpha$

```

>curve(dnorm(x,mean=0,sd=1),from=-3, to=3,axes=FALSE,ylim=c(0,0.5),
+xlax=" ",ylab=" ",main="Densità normale standard")
>text(0,0.05,expression(1-alpha))
>axis(1,c(-3,-1,0,1,3),c(" ",expression(-z[alpha/2]),
+0,expression(z[alpha/2])," "))
>vals<-seq(-3,-1,length=100)
>x<-c(-3,vals,-1,-3)
>y<-c(0,dnorm(vals),0,0)
>polygon(x,y,density=20,angle=45)
>vals<-seq(1,3,length=100)
>x<-c(1,vals,3,1)
>y<-c(0,dnorm(vals),0,0)
>polygon(x,y,density=20,angle=45)
>abline(h=0)
>text(-1.5,0.05,expression(alpha/2))
>text(1.5,0.05,expression(alpha/2))
>box()

```

Per riempire un grafico come quello di Figura 9.1 in R occorre utilizzare la funzione `polygon()` che necessita in input di due vettori di coordinate x e y di un poligono che deve essere necessariamente una figura chiusa. Il codice precedente permette di disegnare prima la coda di sinistra della densità normale standard utilizzando un tratteggio (nel grafico è l'area sottostante la curva normale standard tra -3 e -1). Il vettore `vals` contiene una successione di 100 valori tra i due estremi

e il vettore x è costruito in modo tale che la prima e l'ultima coordinata x dei punti del poligono coincidano. Il vettore y deve invece contenere tutti i punti di ordinata pari alla densità normale standard, esclusi i valori estremi in cui si pone y uguale a 0. Infine, `polygon(x,y,density = 20,angle = 45)` traccia finalmente il poligono riempiendolo di linee inclinate di 45 gradi e equispaziate con densità di 20 per pollice.

Dalla (9.3) si ottiene:

$$P\left(\bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Se poniamo

$$\underline{C}_n = \bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad \overline{C}_n = \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

allora $(\underline{C}_n, \overline{C}_n)$ è un intervallo di confidenza di grado $1 - \alpha$ per μ e le statistiche \underline{C}_n e \overline{C}_n rappresentano rispettivamente il limite inferiore ed il limite superiore di tale intervallo. La lunghezza dell'intervallo di confidenza

$$L(X_1, X_2, \dots, X_n; 1 - \alpha) = \overline{C}_n - \underline{C}_n = 2 z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (9.4)$$

è costante per ogni campione osservato (x_1, x_2, \dots, x_n) . Si nota che la lunghezza dell'intervallo diminuisce al crescere della dimensione n del campione casuale. Inoltre, a valori sempre più piccoli di α , corrispondono lunghezze di intervalli di confidenza sempre più ampi.

Sussiste quindi la seguente proposizione.

Proposizione 9.1 *Sia (x_1, x_2, \dots, x_n) un campione osservato di ampiezza n estratto da una popolazione normale con varianza nota σ^2 . Una stima dell'intervallo di confidenza di grado $1 - \alpha$ per il valore medio μ è*

$$\bar{x}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad (9.5)$$

dove

$$\bar{x}_n = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

denota la media campionaria delle n osservazioni.

Esempio 9.1 Un urbanista è interessato alla superficie media μ delle abitazioni di una certa città. A questo scopo osserva un campione di 50 appartamenti

```
> campnorm<-c(112.6, 118.2, 124.8, 122.1, 137.5, 106.7, 123.7,
+ 127.3, 123.2, 125.1, 120.8, 112.9, 117.0, 128.1, 102.9, 119.1,
+ 127.2, 124.8, 118.0, 131.4, 117.0, 118.2, 125.8, 116.2, 118.5,
+ 120.8, 127.1, 125.0, 131.2, 120.2, 126.0, 119.2, 112.4, 124.6,
+ 117.7, 116.1, 125.3, 115.5, 129.6, 119.1, 130.6, 125.3, 128.7,
+ 134.6, 124.5, 117.2, 126.1, 116.1, 116.0, 125.6)
>
> mean(campnorm)
[1] 121.872
```

e trova che $\bar{x}_{50} = 121.872 m^2$. Supponendo che la popolazione da cui proviene il campione sia normale con deviazione standard nota $\sigma = 8 m^2$, determinare una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la superficie media μ delle abitazioni.

In questo caso $\alpha = 0.05$ e $\alpha/2 = 0.025$. Il valore $z_{\alpha/2} = z_{0.025}$ può essere determinato tramite R. Infatti, osservando la Figura 9.1, facendo uso della (9.5) risulta:

```
> alpha<-1-0.95
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.959964
>
> n<-length(campnorm)
> mean(campnorm)-qnorm(1-alpha/2,mean=0,sd=1)*8/sqrt(n)
[1] 119.6546
> mean(campnorm)+qnorm(1-alpha/2,mean=0,sd=1)*8/sqrt(n)
[1] 124.0894
```

Si nota che $z_{0.025} = 1.96$. La stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la superficie media μ delle abitazioni è quindi (119.7, 124.1). Si nota che la media campionaria \bar{x}_{50} è compresa nell'intervallo. \diamond

Spesso si desidera determinare l'ampiezza del campione n in modo tale da ottenere un intervallo di confidenza di lunghezza minore o uguale ad un valore fissato C avendo stabilito il grado di fiducia $1 - \alpha$. Dalla relazione (9.4) si nota che occorre richiedere

$$L(X_1, X_2, \dots, X_n; 1 - \alpha) = \bar{C}_n - \underline{C}_n = 2 z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq C,$$

da cui segue che

$$n \geq \left(\frac{2\sigma}{C} z_{\alpha/2} \right)^2.$$

Esempio 9.2 Un urbanista è interessato alla superficie media μ delle abitazioni di una certa città. A questo scopo desidera stimare il numero minimo di appartamenti da misurare per ottenere una lunghezza dell'intervallo di confidenza per la superficie media μ delle abitazioni minore o uguale a $3 m^2$. Si supponga che la popolazione da cui proviene il campione sia normale con deviazione standard nota $\sigma = 8 m^2$ e che il grado di fiducia sia $1 - \alpha = 0.95$. Il seguente codice R

```
> sigma<-8
> const<-3
> alpha<-1-0.95
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.959964
>
> ceiling(((2*sigma/const)*qnorm(1-alpha/2,mean=0,sd=1))^2)
[1] 110
```

mostra che l'urbanista deve misurare almeno 110 appartamenti affinché la lunghezza dell'intervallo di confidenza sia minore o uguale a $3 m^2$. \diamond

► (Intervallo di confidenza per μ con varianza non nota)

Denotiamo con

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

la varianza campionaria. Tale statistica gode della seguente proprietà

$$E(S_n^2) = \sigma^2, \quad \text{Var}(S_n) = \frac{1}{n} \left[\mu_4 - \frac{n-3}{n-1} \sigma^4 \right],$$

dove μ_4 denota il momento del quarto ordine e σ^2 la varianza della popolazione da cui è stato estratto il campione.

Si può inoltre dimostrare che la variabile aleatoria

$$Q_n = \frac{(n-1)S_n^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 - \left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \right)^2 \quad (9.6)$$

è distribuita con legge chi-quadrato con $n-1$ gradi di libertà.

Per determinare un intervallo di confidenza di grado $1-\alpha$ per il valore medio μ nel caso in cui la varianza σ^2 della popolazione normale non è nota, utilizziamo il metodo pivotale e consideriamo la variabile aleatoria di pivot.

$$T_n = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}.$$

Tale variabile aleatoria dipende dal campione casuale e dal parametro non noto σ^2 e, quindi, può essere interpretata come una variabile aleatoria di pivot. Inoltre, poichè

$$T_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sqrt{\frac{\sigma^2}{S_n^2}} = \frac{Z_n}{\sqrt{Q_n/(n-1)}},$$

dal Teorema 7.4 segue che T_n è distribuita con legge di Student con $n-1$ gradi di libertà. Scegliendo nel metodo pivotale $\alpha_1 = -t_{\alpha/2, n-1}$ e $\alpha_2 = t_{\alpha/2, n-1}$, dove $t_{\alpha/2, n-1}$ è tale che

$$P(T_n < -t_{\alpha/2, n-1}) = P(T_n > t_{\alpha/2, n-1}) = \frac{\alpha}{2},$$

si ha:

$$P(-t_{\alpha/2, n-1} < T_n < t_{\alpha/2, n-1}) = 1 - \alpha. \quad (9.7)$$

Ciò è evidenziato in Figura 9.2 ottenuta con $n=6$ tramite il seguente codice :

```
>curve(dt(x,df=5),from=-3, to=3,axes=FALSE,ylim=c(0,0.5),xlab="",
+ylab="",main="Densità di Student con n-1 gradi di libertà")
>text(0,0.05,expression(1-alpha))
>axis(1,c(-3,-1,0,1,3),c("",expression(-t[list(alpha/2,n-1)]),0,
+expression(t[list(alpha/2,n-1)]),""))
>vals<-seq(-3,-1,length=100)
>x<-c(-3,vals,-1,-3)
>y<-c(0,dt(vals,df=5),0,0)
```

A.G. Nobile

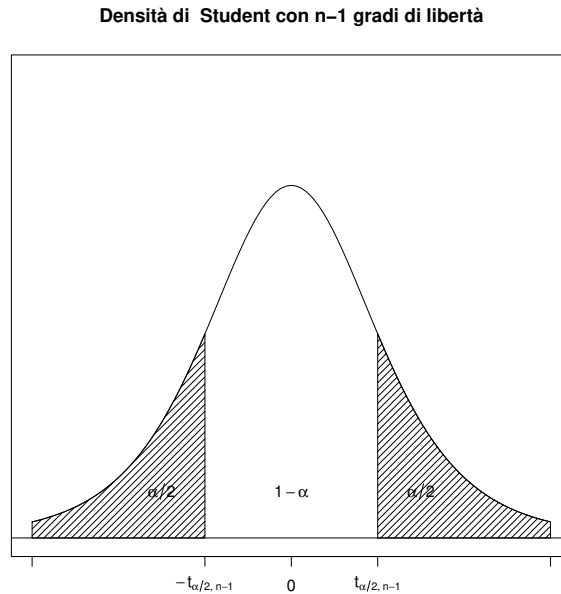


Figura 9.2: Densità di Student con $n-1$ gradi di libertà e grado di fiducia $1-\alpha$

```
>polygon(x,y,density=20,angle=45)
>vals<-seq(1,3,length=100)
>x<-c(1,vals,3,1)
>y<-c(0,dt(vals,df=5),0,0)
>polygon(x,y,density=20,angle=45)
>abline(h=0)
>text(-1.5,0.05,expression(alpha/2))
>text(1.5,0.05,expression(alpha/2))
>box()
```

Nel codice precedente in `expression()` si è usata la funzione `list(x,y)` che fornisce una lista di `x` e `y` separata da virgole; per creare invece una concatenazione di `x` e `y` non separata da virgole si utilizza invece la funzione `paste(x,y)`.

Dalla (9.7) si ottiene:

$$P\left(\bar{X}_n - t_{\alpha/2, n-1} \frac{S_n}{\sqrt{n}} < \mu < \bar{X}_n + t_{\alpha/2, n-1} \frac{S_n}{\sqrt{n}}\right) = 1 - \alpha.$$

Se poniamo

$$\underline{C}_n = \bar{X}_n - t_{\alpha/2, n-1} \frac{S_n}{\sqrt{n}}, \quad \bar{C}_n = \bar{X}_n + t_{\alpha/2, n-1} \frac{S_n}{\sqrt{n}},$$

allora $(\underline{C}_n, \bar{C}_n)$ è un intervallo di confidenza di grado $1-\alpha$ per μ e le statistiche \underline{C}_n e \bar{C}_n rappresentano rispettivamente il limite inferiore ed il limite superiore

di tale intervallo. La lunghezza dell'intervallo di confidenza è

$$L(X_1, X_2, \dots, X_n; 1 - \alpha) = \overline{C}_n - \underline{C}_n = 2 t_{\alpha/2, n-1} \frac{S_n}{\sqrt{n}} \quad (9.8)$$

e si nota che per ogni fissato campione osservato (x_1, x_2, \dots, x_n) essa cresce al diminuire di α . Quindi, per ogni fissato campione osservato (x_1, x_2, \dots, x_n) a valori sempre più piccoli di α (che esprime la probabilità di conclusioni errate), corrispondono lunghezze di intervalli di confidenza sempre più ampi.

Sussiste quindi la seguente proposizione.

Proposizione 9.2 *Sia (x_1, x_2, \dots, x_n) un campione osservato di ampiezza n estratto da una popolazione normale con varianza non nota. Una stima dell'intervallo di confidenza di grado $1 - \alpha$ per il valore medio μ è*

$$\overline{x}_n - t_{\alpha/2, n-1} \frac{s_n}{\sqrt{n}} < \mu < \overline{x}_n + t_{\alpha/2, n-1} \frac{s_n}{\sqrt{n}} \quad (9.9)$$

dove

$$\overline{x}_n = \frac{x_1 + x_2 + \dots + x_n}{n}, \quad s_n = \left\{ \frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x}_n)^2 \right\}^{1/2}$$

denotano rispettivamente la media campionaria e la deviazione standard campionaria delle n osservazioni.

Esempio 9.3 Un produttore di una certa marca di sigarette desidera controllare il quantitativo medio di nicotina in esse contenuto. A questo scopo egli osserva un campione di 30 sigarette

```
> campnorm<-c(10.2, 11.4, 9.7, 10.9, 11.0, 11.3, 9.8, 10.1,
+ 10.8, 10.43, 11.4, 10.8, 11.5, 10.9, 10.0, 11.2, 11.8, 11.8,
+ 10.9, 10.9, 10.9, 11.2, 11.3, 10.6, 10.9, 11.2, 11.5, 11.6,
+ 10.3, 10.8)
>
> mean(campnorm)
[1] 10.90433
> sd(campnorm)
[1] 0.563864
```

e trova che $\overline{x}_{30} = 10.90 \text{ mg}$ e $s_{30} = 0.56 \text{ mg}$. Supponendo che la popolazione da cui proviene il campione sia normale, determinare una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.99$ per il quantitativo medio di nicotina contenuto in una sigaretta.

In questo caso $\alpha = 0.01$ e $\alpha/2 = 0.005$. Il valore $t_{\alpha/2, n-1} = t_{0.005, 29}$ può essere determinato tramite R. Infatti, osservando la Figura 9.2, dalla (9.9) segue che

```
> alpha<-1-0.99
> n<-length(campnorm)
> qt(1-alpha/2, df=n-1)
[1] 2.756386
```

```

> mean(campnorm)-qt(1-alpha/2,df=n-1)*sd(campnorm)/sqrt(n)
[1] 10.62057
> mean(campnorm)+qt(1-alpha/2,df=n-1)*sd(campnorm)/sqrt(n)
[1] 11.1881

```

Si nota che $t_{0.005,29} = 2.756$. La stima dell'intervallo di confidenza di grado $1 - \alpha = 0.99$ per il quantitativo medio di nicotina contenuto in una sigaretta è quindi (10.62, 11.19). \diamond

Esempio 9.4 Ad un campione di 100 studenti che frequentano un certo corso universitario è stato chiesto di assegnare un voto da 1 (pessimo) a 10 (ottimo) come valutazione del corso. La media campionaria del punteggio è risultata $\bar{x}_{100} = 7.2$ con una varianza campionaria $s_{100}^2 = 2.25$. Si desidera

- i)* costruire un intervallo di confidenza del 95% per il punteggio medio assegnato dai 100 studenti;
- ii)* costruire un intervallo di confidenza del 99% per il punteggio medio assegnato dai 100 studenti;
- iii)* se invece di 100 studenti si considerano 200 studenti e risulta $\bar{x}_{200} = 7.2$ e $s_{200}^2 = 2.25$, costruire un intervallo di confidenza del 95% per il punteggio medio assegnato dagli studenti.

Nel caso *i)* si ha $\bar{x}_{100} = 7.2$, $s_{100}^2 = 2.25$ e $1 - \alpha = 0.95$, da cui $\alpha = 0.05$ e $\alpha/2 = 0.025$. Utilizzando R si ha

```

> m<-7.2
> s2<-2.25
> alpha<-1-0.95
> n<-100
> qt(1-alpha/2,df=n-1)
[1] 1.984217
>
> m-qt(1-alpha/2,df=n-1)*sqrt(s2)/sqrt(n)
[1] 6.902367
> m+qt(1-alpha/2,df=n-1)*sqrt(s2)/sqrt(n)
[1] 7.497633

```

Si nota che $t_{0.025,99} = 1.9842$. La stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per il punteggio medio assegnato dagli studenti è quindi (6.902, 7.498).

Nel caso *ii)* si ha $\bar{x}_{100} = 7.2$, $s_{100}^2 = 2.25$ e $1 - \alpha = 0.99$, da cui $\alpha = 0.01$ e $\alpha/2 = 0.005$. Utilizzando R si ha

```

> m<-7.2
> s2<-2.25
> alpha<-1-0.99
> n<-100
> qt(1-alpha/2,df=n-1)
[1] 2.626405
>
> m-qt(1-alpha/2,df=n-1)*sqrt(s2)/sqrt(n)
[1] 6.806039
> m+qt(1-alpha/2,df=n-1)*sqrt(s2)/sqrt(n)
[1] 7.593961

```

Si nota che $t_{0.005,99} = 2.6264$. La stima dell'intervallo di confidenza di grado $1 - \alpha = 0.99$ per il punteggio medio assegnato dagli studenti è quindi (6.806, 7.594). Si nota che aumentando il grado di fiducia aumenta la lunghezza dell'intervallo di confidenza.

Infine, per quanto riguarda il punto *iii*) si ha $\bar{x}_{200} = 7.2$, $s_{200}^2 = 2.25$ e $1 - \alpha = 0.95$, da cui $\alpha = 0.05$ e $\alpha/2 = 0.025$. Utilizzando R si ha:

```
> m<-7.2
> s2<-2.25
> alpha<-1-0.95
> n<-200
> qt(1-alpha/2,df=n-1)
[1] 1.971957
>
> m-qt(1-alpha/2,df=n-1)*sqrt(s2)/sqrt(n)
[1] 6.990842
> m+qt(1-alpha/2,df=n-1)*sqrt(s2)/sqrt(n)
[1] 7.409158
```

Si nota che $t_{0.025,199} = 1.972$. La stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per il punteggio medio assegnato dagli studenti è quindi (6.991, 7.409). Quindi, a parità del livello di fiducia, della media campionaria e della varianza campionaria, all'aumentare della numerosità del campione si riduce l'ampiezza dell'intervallo di confidenza. \diamond

Relativamente all'Esempio 9.4 si potrebbe simulare l'esperimento considerando un campione di voti ottenendo da esso la media campionaria e la varianza campionaria. In R esiste la funzione

```
sample(x, size, replace = FALSE, prob = NULL)
```

dove x è un vettore di valori interi positivi distinti assunti dalla variabile aleatoria discreta X a cui è associato un vettore di probabilità **prob**; **size** è la lunghezza della sequenza di numeri pseudocasuali che simulano X , **replace** indica se le estrazioni sono effettuate con reinserimento (**TRUE**) oppure senza reinserimento (**FALSE**). Se si omette di specificare il vettore **prob** la distribuzione di probabilità di X sarà per default quella equiprobabile.

Ad esempio, per simulare i risultati di 30 prove indipendenti di Bernoulli in cui la probabilità di successo è $p = 1/2$ basta considerare l'istruzione

```
> sample(c(0,1),30,replace=TRUE,prob=c(1/2,1/2))
[1] 0 1 0 1 1 1 0 1 1 1 1 1 1 0 1 0 0 1 1 0 1 0 0 1 0 0
```

mentre per simulare i voti da 1 a 10 assegnati da 30 studenti basta considerare l'istruzione

```
> sample(1:10,30,replace=TRUE)
[1] 4 2 4 7 6 1 4 7 1 9 1 2 3 3 2 3 8 1 5 1 1 6 6 8 9 4 3 3 6 8
```

► (Intervallo di confidenza per σ^2 con μ noto)

Per determinare un intervallo di confidenza di grado $1 - \alpha$ per la varianza nel caso in cui il valore medio μ della popolazione normale è noto, utilizziamo

A.G. Nobile

nuovamente il metodo pivotale e consideriamo la variabile aleatoria di pivot

$$V_n = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$$

Tale variabile dipende dal campione casuale e dal parametro non noto σ^2 (essendo il valore medio μ noto) ed è distribuita con legge chi-quadrato con n gradi di libertà, essendo costituita dalla somma dei quadrati di n variabili aleatorie normali standard. Scegliendo nel metodo pivotale $\alpha_1 = \chi_{1-\alpha/2,n}^2$ e $\alpha_2 = \chi_{\alpha/2,n}^2$ in maniera tale che

$$P(0 < V_n < \chi_{1-\alpha/2,n}^2) = P(V_n > \chi_{\alpha/2,n}^2) = \frac{\alpha}{2}$$

si ha

$$P(\chi_{1-\alpha/2,n}^2 < V_n < \chi_{\alpha/2,n}^2) = 1 - \alpha. \quad (9.10)$$

Ciò è evidenziato in Figura 9.3 ottenuta con $n = 6$ tramite il seguente codice:

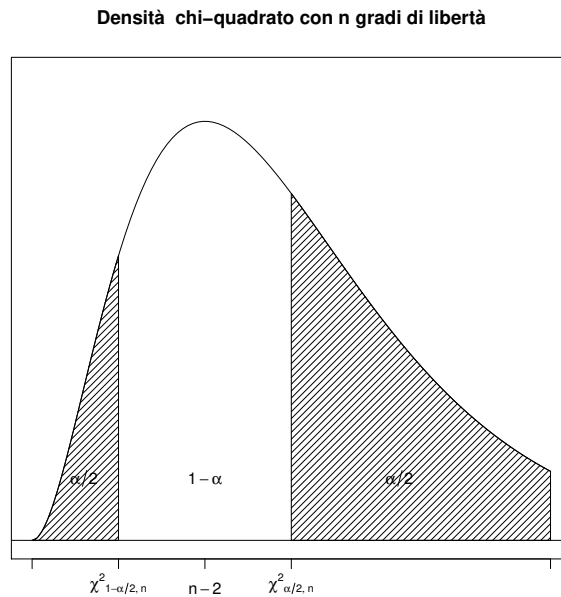


Figura 9.3: Densità chi-quadrato con n gradi di libertà e grado di fiducia $1 - \alpha$

```
>curve(dchisq(x,df=6),from=0, to=12,axes=FALSE,ylim=c(0,0.15),
+      xlab="",ylab="",main="Densità 'chi-quadrato' con n gradi di libertà",
+      )
>text(4,0.02,expression(1-alpha))
>axis(1,c(0,2,4,6,12),c("",expression({chi^2}[list(1-alpha/2,n)]),
```

```

+expression(n-2),expression({chi^2}[list(alpha/2,n)],""))
>vals<-seq(0,2,length=100)
>x<-c(0,vals,2,0)
>y<-c(0,dchisq(vals,df=6),0,0)
>polygon(x,y,density=20,angle=45)
>vals<-seq(6,12,length=100)
>x<-c(6,vals,12,6)
>y<-c(0,dchisq(vals,df=6),0,0)
>polygon(x,y,density=20,angle=45)
>abline(h=0)
>text(1.2,0.02,expression(alpha/2))
>text(8.5,0.02,expression(alpha/2))
>box()

```

Poichè V_n si può scrivere in forma alternativa in termini della media campionaria e della varianza campionaria

$$\begin{aligned}
 V_n &= \sum_{i=1}^n \left(\frac{X_i - \bar{X} + \bar{X} - \mu}{\sigma} \right)^2 = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + \sum_{i=1}^n \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 \\
 &= \frac{(n-1)S_n^2}{\sigma^2} + \left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \right)^2,
 \end{aligned} \tag{9.11}$$

dalla (9.10) si ottiene

$$P\left(\chi_{1-\alpha/2,n}^2 < \frac{(n-1)S_n^2}{\sigma^2} + \left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \right)^2 < \chi_{\alpha/2,n}^2\right) = 1 - \alpha.$$

o equivalentemente

$$P\left(\frac{(n-1)S_n^2 + n(\bar{X}_n - \mu)^2}{\chi_{\alpha/2,n}^2} < \sigma^2 < \frac{(n-1)S_n^2 + n(\bar{X}_n - \mu)^2}{\chi_{1-\alpha/2,n}^2}\right) = 1 - \alpha.$$

Se poniamo

$$\underline{C}_n = \frac{(n-1)S_n^2 + n(\bar{X}_n - \mu)^2}{\chi_{\alpha/2,n}^2}, \quad \overline{C}_n = \frac{(n-1)S_n^2 + n(\bar{X}_n - \mu)^2}{\chi_{1-\alpha/2,n}^2}$$

allora $(\underline{C}_n, \overline{C}_n)$ è un intervallo di confidenza di grado $1 - \alpha$ per σ^2 e le statistiche \underline{C}_n e \overline{C}_n rappresentano rispettivamente il limite inferiore ed il limite superiore di tale intervallo. Abbiamo così dimostrato la seguente proposizione.

Proposizione 9.3 *Sia (x_1, x_2, \dots, x_n) un campione osservato di ampiezza n estratto da una popolazione normale con valore medio noto μ . Una stima dell'intervallo di confidenza di grado $1 - \alpha$ per la varianza σ^2 è*

$$\frac{(n-1)s_n^2 + n(\bar{x}_n - \mu)^2}{\chi_{\alpha/2,n}^2} < \sigma^2 < \frac{(n-1)s_n^2 + n(\bar{x}_n - \mu)^2}{\chi_{1-\alpha/2,n}^2}, \tag{9.12}$$

dove

$$\bar{x}_n = \frac{x_1 + x_2 + \dots + x_n}{n}, \quad s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2.$$

denotano rispettivamente la media campionaria e la varianza campionaria delle n osservazioni.

Esempio 9.5 Osservando un campione contenente il peso in grammi di 12 uova prodotte da un'azienda agricola

```
> campnorm<-c(69.6, 82.2, 64.4, 74.8, 71.2, 70.2, 71.3, 70.6,
+ 72.0, 65.8, 70.3, 63.5)
>
> mean(campnorm)
[1] 70.49167
> var(campnorm)
[1] 24.36447
```

si nota che $\bar{x}_{12} = 70.49 \text{ gr}$ e $s_{12}^2 = 24.36 \text{ gr}^2$. Supponendo che il peso sia distribuito normalmente con valore medio $\mu = 70 \text{ gr}$ e varianza non nota σ^2 , determinare una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la varianza σ^2 .

In questo caso $\alpha = 0.05$ e quindi $\alpha/2 = 0.025$ e $1 - \alpha/2 = 0.975$. I valori $\chi_{1-\alpha/2,n}^2 = \chi_{0.975,12}^2$ possono essere ottenuti tramite R. Facendo riferimento alla Figura 9.3 e ricordando la (9.12) risulta:

```
> n<-length(campnorm)
> mu<-70
>
> alpha<-1-0.95
> qchisq(alpha/2,df=n)
[1] 4.403789
> qchisq(1-alpha/2,df=n)
[1] 23.33666
>
> ((n-1)*var(campnorm)+n*(mean(campnorm)-mu)**2)/qchisq(1-alpha/2,
+ df=n)
[1] 11.60877
> ((n-1)*var(campnorm)+n*(mean(campnorm)-mu)**2)/qchisq(alpha/2,df=
+ n)
[1] 61.51749
```

Si nota che $\chi_{1-\alpha/2,n}^2 = \chi_{0.975,12}^2 = 4.404$ e $\chi_{\alpha/2,n}^2 = \chi_{0.025,12}^2 = 23.337$. La stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la varianza della popolazione normale è quindi (11.61, 61.51). \diamond

Dalla (9.11) segue che la variabile aleatoria $(n-1)S_n^2/\sigma^2$ è distribuita con legge chi-quadrato con $n-1$ gradi di libertà; ciò si rivelerà utile nella determinazione dell'intervallo di confidenza per σ^2 quando il valore medio non è noto.

► **(Intervallo di confidenza per σ^2 con valore medio non noto)**

Per determinare un intervallo di confidenza di grado $1 - \alpha$ per la varianza nel caso in cui il valore medio della popolazione normale non è noto, consideriamo la variabile aleatoria di pivot

$$Q_n = \frac{(n-1)S_n^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Tale variabile aleatoria dipende dal campione casuale e dal parametro non noto σ^2 ed è distribuita con legge chi-quadrato con $n - 1$ gradi di libertà. Scegliendo nel metodo pivotale $\alpha_1 = \chi^2_{1-\alpha/2, n-1}$ e $\alpha_2 = \chi^2_{\alpha/2, n-1}$ in maniera tale che

$$P(0 < Q_n < \chi^2_{1-\alpha/2, n-1}) = P(Q_n > \chi^2_{\alpha/2, n-1}) = \frac{\alpha}{2}$$

si ha

$$P(\chi^2_{1-\alpha/2, n-1} < Q_n < \chi^2_{\alpha/2, n-1}) = 1 - \alpha. \quad (9.13)$$

Ciò è evidenziato in Figura 9.4 ottenuta con $n = 7$ tramite il seguente codice:

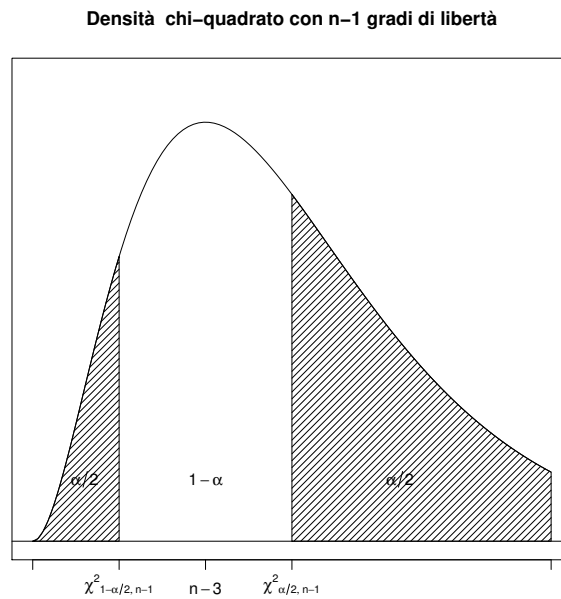


Figura 9.4: Densità chi-quadrato con $n - 1$ gradi di libertà e grado di fiducia $1 - \alpha$

```
>curve(dchisq(x,df=6),from=0, to=12,axes=FALSE,ylim=c(0,0.15),
+xlabs="",ylab="",main="Densità chi-quadrato con n-1
+gradi di libertà")
>text(4,0.02,expression(1-alpha))
>axis(1,c(0,2,4,6,12),c("",expression({chi^2}[list(1-alpha/2,n-1)])
,
+expression(n-2),expression({chi^2}[list(alpha/2,n-1)]),""))
>vals<-seq(0,2,length=100)
>x<-c(0,vals,2,0)
>y<-c(0,dchisq(vals,df=6),0,0)
>polygon(x,y,density=20,angle=45)
>vals<-seq(6,12,length=100)
```

```

>x<-c(6,vals,12,6)
>y<-c(0,dchisq(vals,df=6),0,0)
>polygon(x,y,density=20,angle=45)
>abline(h=0)
>text(1.2,0.02,expression(alpha/2))
>text(8.5,0.02,expression(alpha/2))
>box()

```

Dalla (9.13) si ottiene

$$P\left(\chi_{1-\alpha/2,n-1}^2 < \frac{(n-1)S_n^2}{\sigma^2} < \chi_{\alpha/2,n-1}^2\right) = 1 - \alpha,$$

che è equivalente a richiedere che

$$P\left(\frac{(n-1)S_n^2}{\chi_{\alpha/2,n-1}^2} < \sigma^2 < \frac{(n-1)S_n^2}{\chi_{1-\alpha/2,n-1}^2}\right) = 1 - \alpha.$$

Se poniamo

$$\underline{C}_n = \frac{(n-1)S_n^2}{\chi_{\alpha/2,n-1}^2}, \quad \overline{C}_n = \frac{(n-1)S_n^2}{\chi_{1-\alpha/2,n-1}^2},$$

allora $(\underline{C}_n, \overline{C}_n)$ è un intervallo di confidenza di grado $1 - \alpha$ per σ^2 e le statistiche \underline{C}_n e \overline{C}_n rappresentano rispettivamente il limite inferiore ed il limite superiore di tale intervallo.

Sussiste quindi la seguente proposizione.

Proposizione 9.4 *Sia (x_1, x_2, \dots, x_n) un campione osservato di ampiezza n estratto da una popolazione normale con valore medio non noto. Una stima dell'intervallo di confidenza di grado $1 - \alpha$ per la varianza σ^2 è*

$$\frac{(n-1)s_n^2}{\chi_{\alpha/2,n-1}^2} < \sigma^2 < \frac{(n-1)s_n^2}{\chi_{1-\alpha/2,n-1}^2}, \quad (9.14)$$

dove

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

denota la varianza campionaria delle n osservazioni.

Esempio 9.6 Si supponga che l'errore mensile misurato in secondi commesso da un certo tipo di orologi sia distribuito normalmente con valore medio e varianza non noti. Osservando un campione di 20 orologi

```

> campnorm<-c(-0.47, -0.33, 0.53, -0.32, 0.47, 0.52, 0.21,
+ 0.72, 0.54, -0.06, 0.33, -0.09, 0.37, 0.27, -0.07, -0.51,
+ 0.27, -0.13, -0.04, -0.13)
> mean(campnorm)
[1] 0.104
> var(campnorm)
[1] 0.1339937

```

si nota che $s_{20}^2 = 0.13$. Determinare una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la varianza di una popolazione normale. In questo caso $\alpha = 0.05$ e quindi $\alpha/2 = 0.025$ e $1 - \alpha/2 = 0.975$. I valori $\chi_{1-\alpha/2, n-1}^2 = \chi_{0.975, 19}^2$ e $\chi_{\alpha/2, n-1}^2 = \chi_{0.025, 19}^2$ possono essere ottenuti tramite R. Infatti, osservando la Figura 9.4 e ricordando la (9.14) si ha:

```
> n<-length(campnorm)
> alpha<-1-0.95
>
> qchisq(alpha/2,df=n-1)
[1] 8.906516
> qchisq(1-alpha/2,df=n-1)
[1] 32.85233
>
> (n-1)*var(campnorm)/qchisq(1-alpha/2,df=n-1)
[1] 0.07749466
> (n-1)*var(campnorm)/qchisq(alpha/2,df=n-1)
[1] 0.2858446
```

Si nota che $\chi_{1-\alpha/2, n-1}^2 = \chi_{0.975, 19}^2 = 8.907$ e $\chi_{\alpha/2, n-1}^2 = \chi_{0.025, 19}^2 = 32.852$. La stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la varianza della popolazione normale è quindi $(0.077, 0.286)$. \diamond

Per una popolazione normale le stime per intervallo del valore medio μ e della varianza σ^2 della popolazione possono essere effettuate qualsiasi sia la dimensione del campione casuale osservato. Ciò dipende dalla circostanza favorevole di conoscere la distribuzione esatta della variabile pivotale considerata: normale e di Student per la stima del valore medio e chi-quadrato per la stima della varianza. Occorre anche sottolineare che per una popolazione normale i metodi di stima maggiormente utilizzati sono il (ii) e il (iv), ossia la determinazione di un intervallo di confidenza di grado $1 - \alpha$ per il valore medio μ nel caso in cui la varianza della popolazione normale è non nota e la determinare un intervallo di confidenza di grado $1 - \alpha$ per la varianza σ^2 nel caso in cui il valore medio della popolazione normale è non noto.

