STATISTICA E ANALISI DEI DATI (SECONDA PARTE)

Amelia Giuseppina Nobile[1]

(a.a. 2022/2023)

13 dicembre 2022

[1]Dipartimento 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 $(x_1, x_2, \ldots, x_n)$ 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 $(x_1, x_2, \ldots, x_n)$ 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 $(x_1, x_2, \ldots, x_n) \in A$ si accetta come valida l'affermazione che $\vartheta \in \Theta_0$, mentre se $(x_1, x_2, \ldots, x_n) \in R$ si rifiuta l'ipotesi che $\vartheta \in \Theta_0$ e si accetta l'ipotesi alternativa che $\vartheta \in \Theta_1$. Per affrontare i problemi dell'inferenza statistica, nei prossimi due capitoli introdurremo alcune delle principali variabili aleatorie discrete e continue e le descriveremo con l'ausilio di R.

Variabili aleatorie discrete

Una variabile aleatoria discreta X assume un numero finito o al più numerabile di valori $x_1, x_2, \ldots$ con rispettive probabilità $p_X(x_1), p_X(x_2), \ldots$ essendo $p_X(x_i) = P(X = x_i)$.

Il sistema R mette a disposizione per ciascuna delle principali variabili aleatorie discrete:

- la funzione di probabilità;
- la funzione di distribuzione;
- la funzione per calcolare i quantili;
- la funzione che simula la variabile aleatoria mediante la generazione di sequenze 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 i quantili;

r simula una variabile aleatoria generando una sequenza di numeri pseudocasuali.

I quantili di una variabile aleatoria sono ottenuti utilizzando la definizione di *quantile per una distribuzione di frequenza*, considerata nella statistica descrittiva.

Il quantile (percentile) $z \cdot 100$–esimo di una variabile aleatoria X è definito come il più piccolo numero reale x, assunto dalla variabile aleatoria X, tale che la funzione di distribuzione $F_X(x) = P(X \leq x)$ assuma valori maggiori o uguali di z.

$$F_X(x) = P(X \leq x) \geq z, \qquad 0 \leq z \leq 1.$$

Se si è interessati ai *quartili della variabile aleatoria* basta porre $z = 0.25$ (primo quartile o 25–esimo percentile), $z = 0.5$ (secondo quartile o mediana di una distribuzione di frequenza o 50–esimo percentile), $z = 0.75$ (terzo quartile o 75–esimo percentile).

Variabili aleatorie continue

Una variabile aleatoria continua X *assume un insieme continuo di valori con una densità di probabilità* $f_X(x)$, a differenza delle variabili aleatorie discrete per le quali l'insieme dei possibili valori è finito o al più numerabile.

Il sistema R mette a disposizione per ciascuna delle principali variabili aleatorie 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 (*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 i quantili;

r simula una variabile aleatoria generando una sequenza di numeri pseudocasuali.

Il quantile (percentile) $z \cdot 100$–esimo è definito come il più piccolo numero reale x, assunto dalla variabile aleatoria continua X, tale che

$$F_X(x) = P(X \leq x) \geq z, \qquad 0 \leq z \leq 1.$$

Nei prossimi due capitoli considereremo le seguenti variabili aleatorie:

- *Discrete*: Bernoulli, binomiale, geometrica (geometrica modificata), binomiale negativa (binomiale negativa modificata), Poisson, ipergeometrica;

- *Continue*: uniforme, esponenziale, normale, chi-quadrato, di Student.

Analizzeremo vari tipi di popolazioni, descrivibili con tali variabili aleatorie. Considereremo metodi statistici, utilizzando R, che permettano di estendere le misure ricavate da un campione all'intera popolazione.

Ci occuperemo della stima puntuale e per intervallo dei parametri non noti di una popolazione. Affronteremo poi alcuni problemi di verifica di ipotesi statistiche sui parametri non noti di una popolazione. Infine, utilizzeremo il criterio del chi-quadrato per verificare se il campione osservato può essere stato estratto da una popolazione descritta da una particolare variabile aleatoria.

Capitolo 8

Variabili aleatorie discrete con R

8.1 Introduzione

In questo capitolo considereremo le seguenti distribuzioni discrete:

- distribuzione di Bernoulli;

- distribuzione binomiale;

- distribuzione geometrica;

- distribuzione geometrica modificata;

- distribuzione binomiale negativa;

- distribuzione binomiale negativa modificata;

- distribuzione di Poisson;

- distribuzione ipergeometrica.

8.2 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 $1-p$.

Un tipico esempio è il lancio di una moneta in cui la probabilità di ottenere testa è p. Se codifichiamo testa con 1 e croce con 0, il risultato del lancio della moneta può essere descritto da una variabile aleatoria di Bernoulli con parametro p.

Definizione 8.1 *Una variabile aleatoria X di funzione di probabilità*

$$p_X(x) = P(X = x) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \\ 0, & \textit{altrimenti,} \end{cases} \tag{8.1}$$

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

La distribuzione di Bernoulli prende il nome da un matematico svizzero Jacob Bernoulli (1654–1705) che scoprì non solo tale distribuzione ma anche la distribuzione binomiale. 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*.

La distribuzione di Bernoulli è utilizzata, ad esempio, per modellare

- un sistema informatico che è attivo o inattivo;
- un blocco di dati trasmesso lungo la rete che raggiunge o non raggiunge la destinazione;
- componenti elettroniche funzionanti o difettose;
- parti che superano o falliscono un test;
- hardware funzionante o non funzionante.

Per una variabile aleatoria di Bernoulli si ha:

$$E(X) = p, \qquad E(X^2) = p, \qquad \mathrm{Var}(X) = E(X^2) - [E(X)]^2 = p(1-p).$$

La funzione di distribuzione di X è pertanto

$$F_X(x) = P(X \le x) = \begin{cases} 0, & x < 0 \\ 1-p, & 0 \le x < 1 \\ 1, & x \ge 1. \end{cases} \tag{8.2}$$

Se la popolazione in considerazione è descrivibile mediante una variabile aleatoria di Bernoulli $X \sim \mathcal{B}(1, p)$, nei prossimi capitoli affronteremo i problemi di stimare il valore medio $E(X) = p$ e di effettuare alcuni test di verifica di ipotesi sul parametro p utilizzando un campione casuale estratto dalla stessa popolazione.

8.3 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 $\overline{A}$ (interpretabile come insuccesso), con $P(A) = p$ $(0 < p < 1)$. Un siffatto esperimento si dice costituito da *n prove ripetute indipendenti di Bernoulli*.

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

Definizione 8.2 *Una variabile aleatoria X di funzione di probabilità*

$$p_X(x) = P(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, \ldots, n \\ 0, & \textit{altrimenti,} \end{cases} \tag{8.3}$$

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

Il termine *binomiale* deriva dalla contrazione di *bi* (che significa *due*) e *nom* (che significa *un numero*), riflettendo così il concetto di risultati binari.

La distribuzione binomiale è utile, ad esempio, per modellare

- il numero di successi in una sequenza di lanci indipendenti di una moneta;
- il numero di processori che sono attivi in un sistema multiprocessore;
- il numero di blocchi di dati trasmessi lungo la rete che raggiunge la destinazione senza perdite;
- il numero di computer difettosi in una spedizione;
- il numero di file aggiornati in una cartella;
- il numero di e-mail con allegati;
- il numero di articoli in un lotto che hanno determinate caratteristiche.

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 (8.3) si riduce alla funzione di probabilità di Bernoulli di parametro p. Dalla (8.3) si ricava:

$$\frac{p_X(x)}{p_X(x-1)} = \frac{p}{1-p}\,\frac{n-x+1}{x}, \qquad x = 1, 2, \ldots, n, \tag{8.4}$$

da cui segue che le probabilità binomiali (8.3) sono calcolabili ricorsivamente al seguente modo:

$$p_X(0) = (1-p)^n, \qquad p_X(x) = \frac{p}{1-p}\,\frac{n-x+1}{x}\,p_X(x-1), \qquad x = 1, 2, \ldots, n.$$

La funzione di distribuzione di X è poi immediatamente ottenibile:

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \displaystyle\sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}, & k \le x < k+1 \qquad (k = 0, 1, \ldots, n-1) \\ 1, & x \ge n. \end{cases} \tag{8.5}$$

Per una variabile aleatoria binomiale si ha

$$X = X_1 + X_2 + \ldots + X_n,$$

dove $X_1, X_2, \ldots, X_n$ sono variabili aleatorie di Bernoulli indipendenti ed identicamente distribuite. Pertanto, per una variabile aleatoria binomiale si ottiene:

$$E(X) = n\,p, \qquad \mathrm{Var}(X) = n\,p\,(1-p). \tag{8.6}$$

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

Per il calcolo in R delle probabilità binomiali si utilizza la funzione:

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

dove

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

Le seguenti linee di codice permettono di visualizzare le funzioni di probabilità binomiali di Figura 8.1.

```
> 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")
```

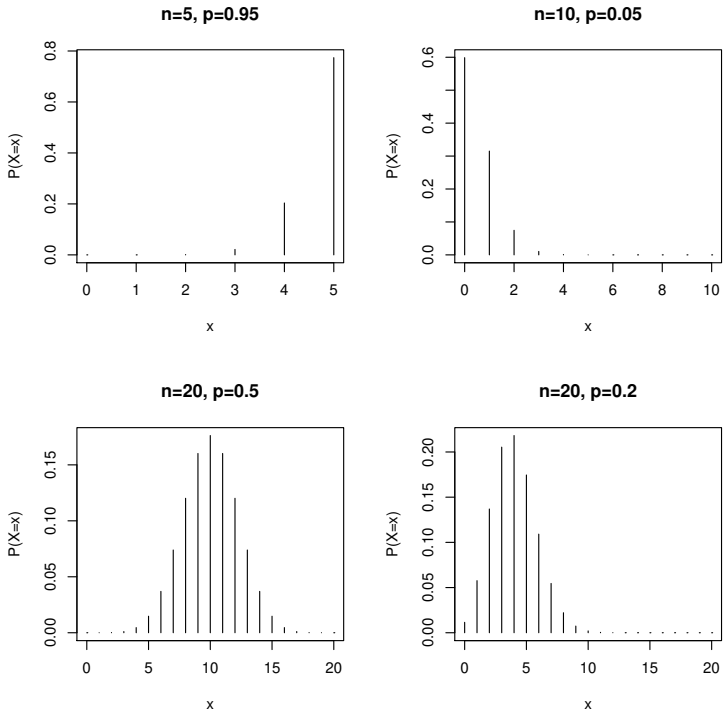

Figura 8.1: Funzione di probabilità binomiale per alcuni valori di n e p.

Per il calcolo della funzione di distribuzione binomiale in R si utilizza la funzione:

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

dove

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria binomiale;
- 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 $P(X > x)$.

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

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

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:

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

Le seguenti linee di codice permettono di visualizzare le funzioni di distribuzione di Figura 8.2.

```
> 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")
```

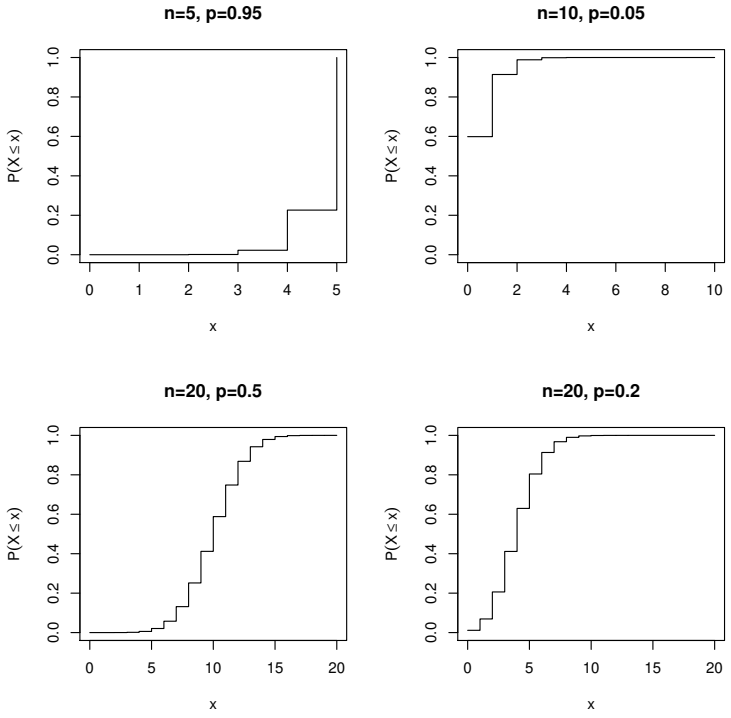

Figura 8.2: Funzione di distribuzione binomiale per alcuni valori di n e p.

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
```

mostrano che $E(X) = 4$, $\mathrm{Var}(X) = 3.2$, $\sqrt{\mathrm{Var}(X)} = 1.7888544$ e $\mathrm{CV}(X) = 0.4472136$. Si può utilizzare direttamente la (8.6) avendosi $E(X) = n\,p = 20 \cdot 0.2 = 4$ e $\mathrm{Var}(X) = n\,p\,(1-p) = 20 \cdot 0.2 \cdot 0.8 = 3.2$.

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

```
qbinom(z, size, prob)
```

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;

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*

$$F_X(x) = P(X \leq k) \geq z \qquad (k = 0, 1, \ldots, n). \tag{8.7}$$

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 $Q_4 = 20$.

Esempio 8.1 *(Canale di comunicazione binario)* Consideriamo un canale di comunicazione binario che trasmette parole di codice di n bits. Assumiamo che la probabilità di successo nella trasmissione di un singolo bit sia p e che la probabilità di un errore sia $q = 1 - p$. Se assumiamo che la trasmissione di bit successivi avviene indipendentemente, siamo interessati a calcolare la probabilità di commettere al più n_e errori nella trasmissione di una parola di codice. Se denotiamo con N la variabile aleatoria che descrive il numero di errori commessi nella trasmissione di una parola di codice, la probabilità di commettere k errori è

$$P(N = k) = \binom{n}{k} q^k p^{n-k}, \qquad k = 0, 1, \ldots, n,$$

e quindi la probabilità di commettere al più n_e errori è:

$$P(N \leq n_e) = \sum_{k=0}^{n_e} \binom{n}{k} q^k p^{n-k} \qquad n_e = 0, 1, \ldots, n.$$

Per calcolare tale probabilità in R basta utilizzare la funzione pbinom(n$_e$, n, q). Ad esempio la probabilità di commettere al più un errore in una parola di codice di lunghezza 10 con probabilità di errore $q = 0.1$ è:

```
> pbinom(1,10,0.1)
[1] 0.7360989
```

$\Diamond$

Esempio 8.2 *(Regola di decisione a maggioranza)* Supponiamo di effettuare n lanci ($n = 3, 5, \ldots$) indipendenti di una moneta e supponiamo che sia p la probabilità di successo e $1 - p$ la probabilità di insuccesso in ogni singola prova. Sia X la variabile aleatoria che descrive il numero di successi in n prove. 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 indipendenti, X ha distribuzione binomiale di parametri n e p; per $n = 3, 5, \ldots$ si ha:

$$Q_n = P(X > n/2) = P\Big(X \geq \frac{n+1}{2}\Big) = \sum_{k=(n+1)/2}^{n} \binom{n}{k} p^k (1-p)^{n-k}.$$

Si può mostrare graficamente con R che se $p > 0.5$ e $n = 3, 5, \ldots$ la probabilità Q_n è una funzione crescente in n. Invece, se $p < 0.5$ e $n = 3, 5, \ldots$ la probabilità Q_n è una funzione decrescente in n. Infine, se $p = 0.5$ risulta $Q_n = 1/2$ per $n = 3, 5, \ldots$

In R, il codice per calcolare la probabilità Q_n per $p = 0.7$ e $n = 3, 5, \ldots, 35$ è

```
> n<-seq(3,35,2)
> Qn<-pbinom(n/2,n,0.7,lower.tail=FALSE)
> 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",main="p=0.7")
```

Le probabilità Q_n per $p = 0.7$ sono rappresentate in Figura 8.3; si nota che Q_n aumenta con il numero di prove e presenta una rapida crescita fino a circa 21 prove, dopo di che la crescita di Q_n diventa molto più lenta.

Il codice per calcolare la probabilità Q_n per $p = 0.3$ e $n = 3, 5, \ldots, 35$ è

```
> n<-seq(3,35,2)
> Qn<-pbinom(n/2,n,0.3,lower.tail=FALSE)
> round(Qn,4)
 [1] 0.2160 0.1631 0.1260 0.0988 0.0782 0.0624 0.0500 0.0403 0.0326
[10] 0.0264 0.0214 0.0175 0.0143 0.0117 0.0095 0.0078 0.0064
> plot(n,Qn,xlab="Numero di prove",
+ ylab=expression('Q'['n']),
+ ylim=c(0,0.25),type="h",main="p=0.3")
```

Le probabilità Q_n per $p = 0.3$ sono rappresentate in Figura 8.4; si nota che Q_n diminuisce con il numero di prove. $\Diamond$

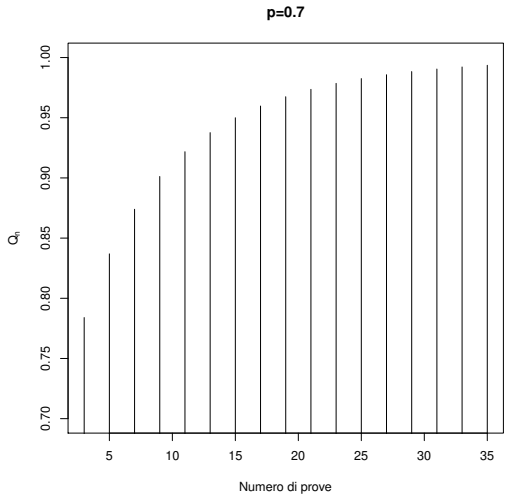

Figura 8.3: Probabilità Q_n per $p = 0.7$ e $n = 3, 5 \ldots, 35$.

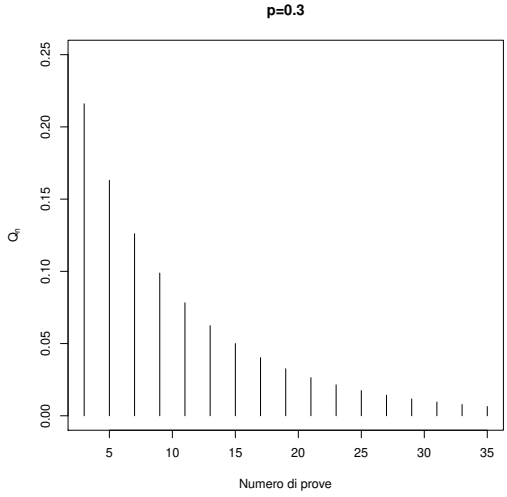

Figura 8.4: Probabilità Q_n per $p = 0.3$ e $n = 3, 5 \ldots, 35$.

È possibile simulare in R la variabile aleatoria binomiale[1] 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) # frequenze assoluta
sim
 1  2  3  4  5  6  7  8
 8  4 12  9  7  4  4  2
> table(sim)/length(sim) # frequenze relative
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 $0, 1, \ldots,$ 20 si presentano nella sequenza generata. Occorre sottolineare che differenti esecuzioni conducono a sequenze pseudocasuali diverse. Per ottenere la stessa sequenza occorre fissare prima di rbinom(N, size, prob) il seme iniziale del generatore tramite la funzione set.seed(seme)

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="Probabilita' ",
+ 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)
```

[1] V. Kachitvichyanukul, B. W. Schmeiser. Binomial random variate generation. Communications of the ACM, 31, 216–222 (1988)

```
> 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))
```

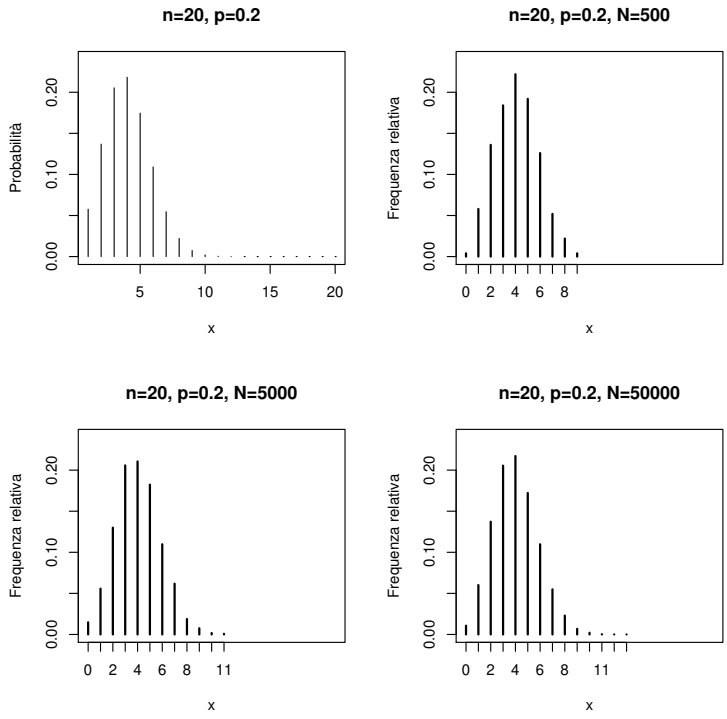

Figura 8.5: Confronto della funzione di probabilità binomiale teorica e delle frequenze relative simulate per una variabile aleatoria binomiale $X \sim \mathcal{B}(20, 0.2)$.

In Figura 8.5 si nota che all'aumentare della lunghezza N della sequenza generata il grafico delle frequenze relative si avvicina sempre di più al grafico della funzione di probabilità binomiale.

Se la popolazione in considerazione è descrivibile mediante una variabile aleatoria binomiale $X \sim \mathcal{B}(n, p)$, nei prossimi capitoli affronteremo i problemi di stimare il valore medio $E(X) = np$ e di effettuare alcuni test di verifica di ipotesi sul parametro p utilizzando un campione casuale estratto dalla stessa popolazione.

8.4 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

$$F_r = \{\textit{il numero di fallimenti che precedono il primo successo è } r\} \quad (r = 0, 1, \ldots).$$

Dall'ipotesi di indipendenza delle prove si ricava che $P(F_r) = (1-p)^r\, p$.

Sia Y la variabile aleatoria che descrive il *numero di fallimenti che precedono il primo successo*; è evidente che $P(Y = r) = P(F_r)$ per $r = 0, 1, \ldots$

Definizione 8.3 *Una variabile aleatoria Y di funzione di probabilità*

$$p_Y(y) = P(Y = y) = \begin{cases} (1-p)^y\, p, & y = 0, 1, \ldots \\ 0, & \textit{altrimenti,} \end{cases} \tag{8.8}$$

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

La distribuzione geometrica è utilizzata, ad esempio, per modellare:

- numero di ritrasmissioni di un messaggio in un sistema informatico;
- numero di fallimenti di uno sportivo per riuscire a completare un percorso senza alcun incidente;
- numero di fallimenti ad una prova di esame prima di superarla;
- numero di tentativi falliti con le chiavi da un ubriaco prima che riesca ad aprire la porta di casa.

Dalla (8.8) segue immediatamente che $p_Y(y)$ è strettamente decrescente in $y = 0, 1, \ldots$. Poiché

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

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

$$F_Y(y) = P(Y \le y) = \begin{cases} 0, & y < 0 \\ 1 - (1-p)^{k+1}, & k \le y < k+1 \quad (k = 0, 1, \ldots). \end{cases} \tag{8.9}$$

Per una variabile aleatoria geometrica Y si ha:

$$E(Y) = \frac{1-p}{p}, \qquad \mathrm{Var}(Y) = \frac{1-p}{p^2}. \tag{8.10}$$

Una proprietà della distribuzione geometrica è la seguente:

$$P(Y > r + n \,|\, Y > r) = P(Y > n), \tag{8.11}$$

con r e n interi non negativi. La (8.11) esprime dunque la proprietà di *assenza di memoria della distribuzione geometrica*, ossia il numero di fallimenti fino al primo successo non dipende da r, ossia da quanto si è atteso, ma solo dal numero n di prove ancora da effettuarsi.

Per il calcolo in R delle probabilità di una variabile aleatoria geometrica Y si utilizza la funzione:

```
dgeom(x, prob)
```

dove

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

- prob è la probabilità di successo in ciascuna prova.

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

```
> 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
```

Le seguenti linee di codice permettono di visualizzare le funzioni di probabilità di una variabile aleatoria geometrica Y come illustrato in Figura 8.6.

```
> 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")
```

È possibile calcolare il valore medio, la varianza, la deviazione standard e il coefficiente di variazione della distribuzione geometrica attraverso la (8.10). Ad esempio, se $p = 0.2$ si ricava $E(Y) = (1-p)/p = 4.0$, $\text{Var}(Y) = (1-p)/p^2 = 20$ e $\text{CV}(Y) = \sqrt{(20)}/4 = 1.118034$.

Per il calcolo della funzione di distribuzione di una variabile aleatoria geometrica Y si utilizza la funzione:

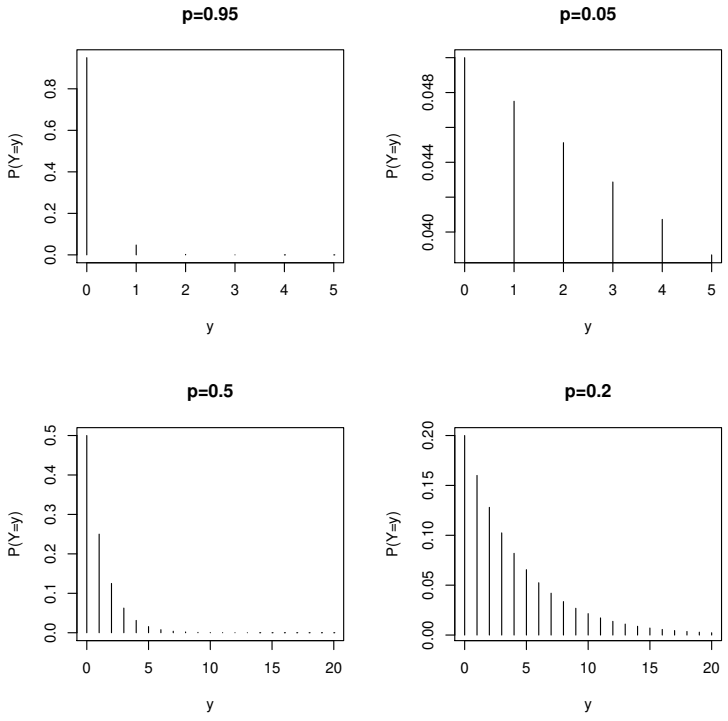

Figura 8.6: Funzione di probabilità geometrica Y per alcuni valori di p.

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

dove

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria geometrica 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 $p = 0.95$ la funzione di distribuzione di una variabile aleatoria geometrica può essere così valutata:

```
> 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

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

Le seguenti linee di codice permettono di visualizzare le funzioni di distribuzione delle variabili aleatorie geometriche di Figura 8.7.

```
> 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")
```

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

```
qgeom(z, prob)
```

dove

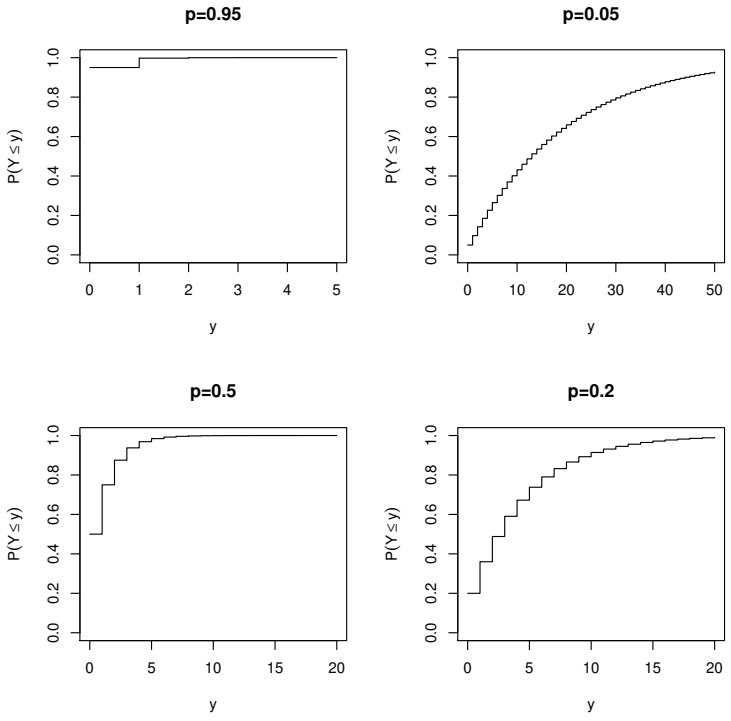

Figura 8.7: Funzione di distribuzione geometrica per alcuni valori di p.

- 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.

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

$$P(Y \leq k) = 1 - (1-p)^{k+1} \geq z \qquad (k = 0, 1, \ldots). \tag{8.12}$$

da cui segue

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

ossia

$$(k+1)\log(1-p) \leq \log(1-z) \qquad (k = 0, 1, \ldots).$$

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

$$k \geq \frac{\log(1-z)}{\log(1-p)} - 1 \qquad (k = 0, 1, \ldots).$$

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

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

In R il risultato della funzione qgeom() è il percentile $z \cdot 100$–esimo della distribuzione geometrica. 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 la variabile aleatoria geometrica[2] in R generando una sequenza di numeri pseudocasuali mediante la funzione

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

[2]L. Devroye. Non-Uniform Random Variate Generation. Springer-Verlag, New York (1986)

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 geometrica 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) # frequenze assolute
sim
 0  1  2  4  5  6  7 12
 5  7  2  2  1  1  1  1
> table(sim)/length(sim) # frequenze relative
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 $0, 1, \ldots$ si presentano nella sequenza generata. Differenti esecuzioni possono condurre a diverse sequenze pseudocasuali.

Se la popolazione in considerazione è descrivibile mediante una variabile aleatoria geometrica Y, nei prossimi capitoli affronteremo i problemi di stimare il valore medio $E(Y) = (1-p)/p$ e di effettuare opportuni test di verifica di ipotesi sul parametro p utilizzando un campione casuale estratto dalla stessa popolazione.

8.5 Distribuzione geometrica modificata

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

$$E_r = \{\textit{il primo successo si verifica alla prova } r\textit{-esima}\} \qquad (r = 1, 2, \ldots).$$

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 $r = 1, 2, \ldots$

Definizione 8.4 *Una variabile aleatoria X di funzione di probabilità*

$$p_X(x) = P(X = x) = \begin{cases} p\,(1-p)^{x-1}, & x = 1, 2, \ldots \\ 0, & \textit{altrimenti,} \end{cases} \tag{8.13}$$

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

Alcune situazioni descrivibili con una distribuzione geometrica sono le seguenti:

⋆ Un motore di ricerca passa in rassegna un elenco di siti alla ricerca di una determinata frase chiave. Supponiamo che la ricerca termini non appena viene trovata la frase chiave. Il numero di siti visitati è descrivibile con una distribuzione geometrica modificata.

⋆ Un responsabile delle assunzioni intervista i candidati, uno per uno, per coprire un posto vacante. Il numero di candidati intervistati fino a quando un candidato riceve un'offerta è descrivibile con una distribuzione geometrica modificata.

La distribuzione geometrica modificata è anche utilizzata per modellare:

- numero di analisi da effettuare in un laboratorio prima di ottenere una risposta positiva;

- numero di farmaci da sperimentare prima di trovarne uno efficace.

Dalla (8.13) segue immediatamente che $p_X(x)$ è strettamente decrescente in $x = 1, 2, \ldots$. Inoltre, sussiste la relazione $X = Y + 1$, dove Y è la variabile aleatoria geometrica; quindi

$$p_X(x) = P(X = x) = P(Y + 1 = x) = P(Y = x - 1) = p_Y(x - 1), \quad x = 1, 2, \ldots$$
$$F_X(x) = P(X \leq x) = P(Y \leq x - 1) = F_Y(x - 1), \qquad x \in \mathbb{R}.$$

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:

$$F_X(x) = \begin{cases} 0, & x < 1 \\ 1 - (1-p)^k, & k \leq x < k+1 \qquad (k = 1, 2, \ldots). \end{cases} \tag{8.14}$$

Per una variabile aleatoria geometrica modificata si ha:

$$E(X) = E(Y + 1) = \frac{1}{p}, \qquad \mathrm{Var}(X) = \mathrm{Var}(Y + 1) = \frac{1-p}{p^2}\cdot \tag{8.15}$$

Per calcolare la funzione di probabilità e la funzione di distribuzione di una variabile aleatoria geometrica modificata si considerano le funzioni:

```
dgeom(x-1, prob)
pgeom(x-1, prob, lower.tail = TRUE)
```

Per visualizzare le funzioni di probabilità di una variabile aleatoria geometrica modificata $X = Y + 1$ con le stesse scelte delle probabilità di Figura 8.6 occorre procedere come segue (vedi Figura 8.8)

```
> par(mfrow=c(2,2))
> x<-1:6
> plot(x,dgeom(x-1,prob=0.95),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="p=0.95")
>
> x<-1:11
> plot(x,dgeom(x-1,prob=0.05),
```

```
+ xlab="x",ylab="P(X=x)",type="h",
+ main="p=0.05")
>
> x<-1:21
> plot(x,dgeom(x-1,prob=0.5),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="p=0.5")
>
> y<-1:21
> plot(x,dgeom(x-1,prob=0.2),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="p=0.2")
```

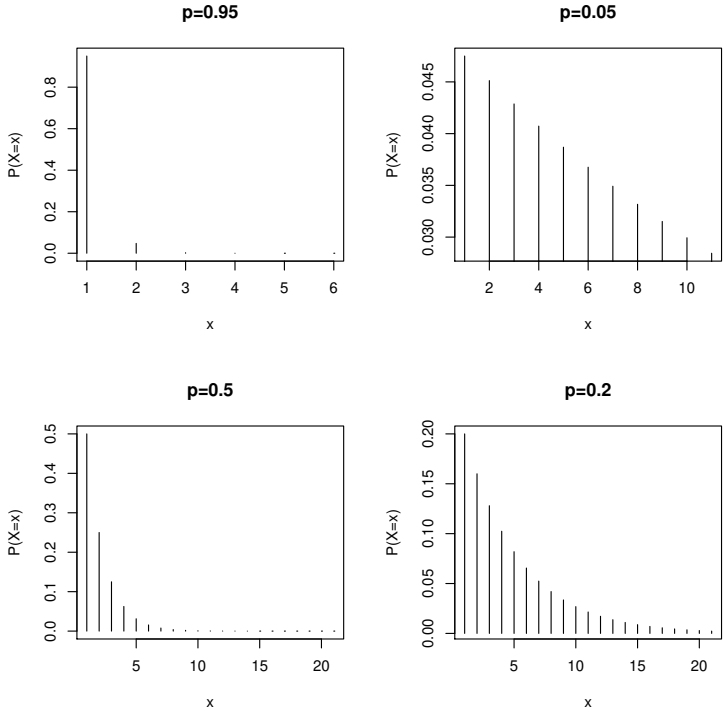

Figura 8.8: Funzione di probabilità geometrica modificata per alcuni valori di p.

Possiamo procedere allo stesso modo per visualizzare la funzione di distribuzione. Per calcolare i quantili oppure per simulare una variabile aleatoria geometrica modificata X, basta ricordare che $X = Y + 1$ e aggiungere 1 ai quantili oppure alla simulazione della variabile geometrica. Ad esempio, se desideriamo generare una sequenza di 20 numeri pseudocasuali simulando una variabile aleatoria geometrica modificata con $p = 0.2$ si ha:

```
> sim<-rgeom(20,prob=0.2)+1
> sim
 [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 $0, 1, \ldots,$ 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 modificata con quella simulata all'aumentare della lunghezza $N = 500, 5000, 50000$ della sequenza generata.

```
> par(mfrow=c(2,2))
> x<-1:21
> plot(x,dgeom(x-1,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 all'aumentare della lunghezza della sequenza generata il grafico delle frequenze relative si avvicina sempre di più al grafico della funzione di probabilità geometrica modificata.

Esempio 8.3 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 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

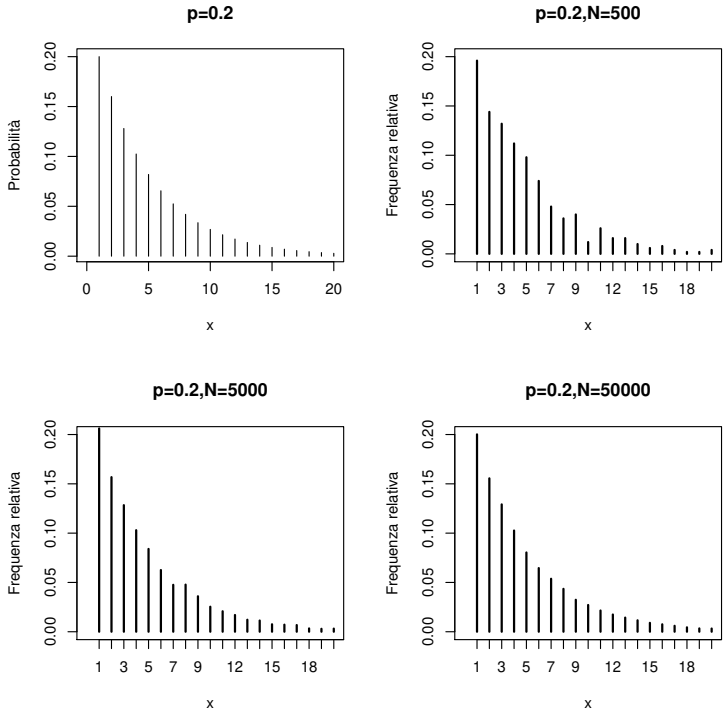

Figura 8.9: Confronto della funzione di probabilità geometrica modificata teorica e delle frequenze relative simulate per una variabile aleatoria geometrica modificata di parametro $p = 0.2$.

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

$$E(X) = \frac{p}{1-p}\Big[\frac{1}{1-2(1-p)} - 1\Big] = \frac{p}{1-p}\,\frac{2(1-p)}{2p-1} = \frac{2p}{2p-1} \quad (p > 1/2)\cdot$$

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

$$\overline{x} = \frac{2^{x_1} + 2^{x_2} + \ldots + 2^{x_N}}{N}.$$

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, è 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 8.10 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.

Se la popolazione in considerazione è descrivibile mediante una variabile aleatoria geometrica modificata X, nei prossimi capitoli affronteremo i problemi di stimare il valore medio $E(X) = 1/p$ e di effettuare opportuni test di verifica di ipotesi sul parametro p utilizzando un campione casuale estratto dalla stessa popolazione.

8.6 Distribuzione binomiale negativa

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

$$F_r = \{\textit{il numero di fallimenti che precedono il successo } n\textit{-esimo in una sequenza di prove di Bernoulli è } r\} \qquad (r = 0, 1, \ldots).$$

Sia Y la variabile aleatoria che descrive il *numero di fallimenti che precedono il successo n-esimo*; è evidente che $P(Y = r) = P(F_r)$ per $r = 0, 1, \ldots$

Definizione 8.5 *Una variabile aleatoria Y di funzione di probabilità*

$$p_Y(y) = P(Y = y) = \begin{cases} \binom{n+y-1}{y} p^n (1-p)^y, & y = 0, 1, \ldots \\ 0, & \textit{altrimenti} \end{cases} \tag{8.16}$$

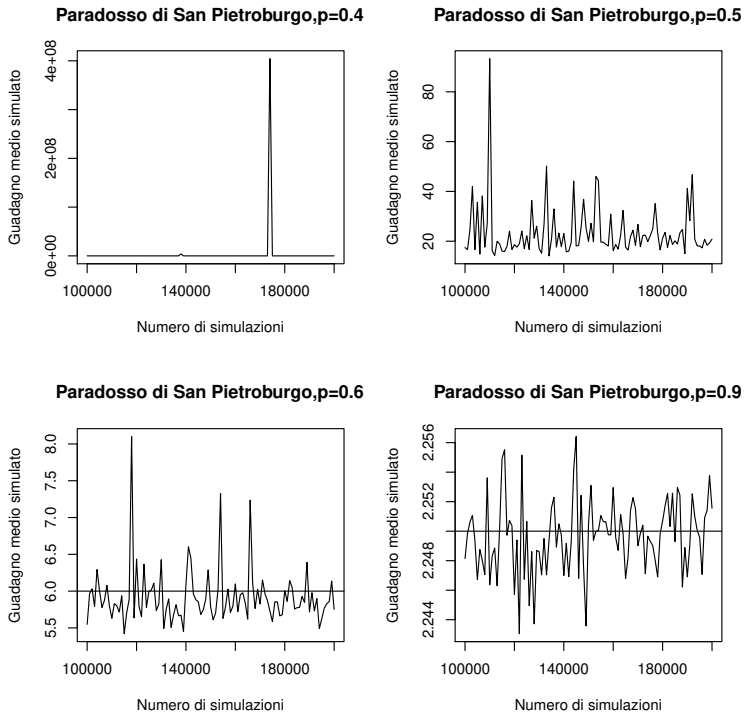

Figura 8.10: Guadagno medio simulato per diverse scelte di p. La linea orizzontale indica il guadagno medio teorico nel caso in cui $p > 0.5$.

con $0 < p < 1$ *si dice avere distribuzione binomiale negativa di parametri* n *e* p*.*

La distribuzione binomiale negativa è utile, ad esempio, per modellare:

- numero di ritrasmissioni di un messaggio costituito da n blocchi in un sistema informatico;
- numero di bit ricevuti senza errori in un collegamento con rumore prima dell'n-esimo bit di errore.

Per la mancanza di memoria della distribuzione geometrica, si ha

$$Y = Y_1 + Y_2 + \ldots + Y_n,$$

dove $Y_1, Y_2, \ldots, Y_n$ sono variabili aleatorie indipendenti di tipo geometrico, descriventi il numero di fallimenti fino al primo successo. Pertanto, risulta che

$$E(Y) = \frac{n\,(1-p)}{p}, \qquad \mathrm{Var}(Y) = \frac{n(1-p)}{p^2}.$$

che fornisce il numero medio e la varianza del numero di fallimenti fino al successo n-esimo.

In R per calcolare la funzione di probabilità binomiale negativa si utilizza la funzione

```
dnbinom(x, size, prob)
```

dove

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria binomiale negativa;
- size è il numero richiesto di successi;
- prob è la probabilità di successo in ciascuna prova.

Per calcolare la funzione di distribuzione binomiale negativa si utilizza la funzione:

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

dove

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria binomiale;
- size è il numero richiesto di successi;
- 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 $P(X > x)$.

Per calcolare i quantili (percentili) della distribuzione binomiale negativa si utilizza la funzione

```
qnbinom(z, size,prob)
```

dove

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

- size è il numero richiesto di successi;

- prob è la probabilità di successo in ciascuna prova.

Infine, per simulare una sequenza di N numeri con distribuzione binomiale negativa si utilizza la funzione:

```
rnbinom(N, size, prob )
```

Ad esempio, generiamo una sequenza di 10 numeri pseudocasuali che rappresentano il numero di fallimenti che si sono verificati fino al secondo successo supponendo che $p = 0.2$:

```
> sim<-rnbinom(10,size=2,prob=0.2)
> sim
 [1]  7  8  3  6  8 12  1  6 13  2
> table(sim)
sim
 1  2  3  6  7  8 12 13
 1  1  1  2  1  2  1  1
> table(sim)/length(sim)
sim
  1   2   3   6   7   8  12  13
0.1 0.1 0.1 0.2 0.1 0.2 0.1 0.1
```

Se la popolazione in considerazione è descrivibile mediante una variabile aleatoria binomiale negativa Y, nei prossimi capitoli affronteremo i problemi di stimare il valore medio $E(Y) = n(1-p)/p$ e di effettuare opportuni test di verifica di ipotesi sul parametro p utilizzando un campione casuale estratto dalla stessa popolazione.

8.7 Distribuzione binomiale negativa modificata

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

$$E_r = \{\textit{il successo } n\textit{-esimo si verifica alla prova } r\textit{-esima in una sequenza di prove di Bernoulli}\} \qquad (r = 1, 2, \ldots).$$

Sia X la variabile aleatoria che descrive il *numero di prove indipendenti necessarie per ottenere il successo n-esimo*; è evidente che $P(X = r) = P(E_r)$ per $r = n, n + 1, \ldots$

Definizione 8.6 *Una variabile aleatoria X di funzione di probabilità*

$$p_X(x) = P(X = x) = \begin{cases} \binom{x-1}{n-1} p^n (1-p)^{x-n}, & x = n, n+1, \dots \\ 0, & \textit{altrimenti} \end{cases} \tag{8.17}$$

con $0 < p < 1$ si dice avere distribuzione binomiale negativa modificata di parametri n e p.

La distribuzione binomiale negativa modificata è utile, ad esempio, per modellare:

- numero di analisi da effettuare in un laboratorio prima di ottenere l'n-esima risposta positiva;
- numero di farmaci da sperimentare prima di trovarne l'n-esimo efficace.

Per la mancanza di memoria della distribuzione geometrica modificata, si ha

$$X = X_1 + X_2 + \ldots + X_n,$$

dove $X_1, X_2, \ldots, X_n$ sono variabili aleatorie indipendenti geometriche modificate, descriventi il numero di prove necessarie per ottenere il primo successo. Pertanto, $X = Y + n$, dove Y è una variabile binomiale negativa; quindi si ottiene:

$$E(X) = E(Y + n) = \frac{n}{p}, \qquad \mathrm{Var}(X) = \mathrm{Var}(Y + n) = \frac{n(1-p)}{p^2}.$$

che fornisce il numero medio e la varianza del numero di prove fino al successo n-esimo. Inoltre, si ha

$$p_X(x) = P(X = x) = P(Y + n = x) = P(Y = x - n) = p_Y(x - n), \quad x = n, n+1,$$
$$F_X(x) = P(X \le x) = P(Y \le x - n) = F_Y(x - n), \quad x \in \mathbb{R}.$$

In R per calcolare la funzione di probabilità binomiale negativa modificata e la funzione di distribuzione si utilizzano le funzioni

```
dnbinom(x-n, size, prob)
pnbinom(x-n, size, prob, lower.tail = TRUE)
```

dove

- x − n è il valore assunto (o i valori assunti) dalla variabile aleatoria binomiale negativa modificata;
- size è il numero richiesto di successi;
- prob è la probabilità di successo in ciascuna prova,
- lower.tail se tale parametro è TRUE (caso di default) calcola $P(X \le x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

Per calcolare i quantili (percentili) della distribuzione binomiale negativa modificata e per simulare tale variabile aleatoria basta aggiungere n ai quantili e alla simulazione della variabile binomiale negativa:

```
qnbinom(z, size,prob)+n
rnbinom(N, size, prob )+n
```

Ad esempio, generiamo una sequenza di 10 numeri pseudocasuali che rappresentano il numero di prove che occorrono per ottenere il secondo successo supponendo che $p = 0.2$.

```
> sim<-rnbinom(10,size=2,prob=0.2)+2
> sim
 [1] 10  5 18 18  5 15 16  8  6  7
> table(sim)
sim
 5  6  7  8 10 15 16 18
 2  1  1  1  1  1  1  2
> table(sim)/length(sim)
sim
  5   6   7   8  10  15  16  18
0.2 0.1 0.1 0.1 0.1 0.1 0.1 0.2
```

Se la popolazione in considerazione è descrivibile mediante una variabile aleatoria binomiale negativa modificata X, nei prossimi capitoli affronteremo i problemi di stimare il valore medio $E(X) = n/p$ e di effettuare opportuni test di verifica di ipotesi sul parametro p utilizzando un campione casuale estratto dalla stessa popolazione.

8.8 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 8.7 *Una variabile aleatoria X avente funzione di probabilità*

$$p_X(x) = P(X = x) = \begin{cases} \dfrac{\lambda^x}{x!}\, e^{-\lambda}, & x = 0, 1, \ldots \qquad (\lambda > 0), \\ 0, & \textit{altrimenti} \end{cases} \tag{8.18}$$

è detta di distribuzione di Poisson di parametro λ.

Questa distribuzione porta il nome di un famoso matematico francese Siméon–Denis Poisson (1781-1840). 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*.

La distribuzione di Poisson è connessa a un concetto di *eventi rari* o *eventi poissoniani* (è improbabile che si verifichi più di un evento in un piccolo intervallo di tempo). Le telefonate in arrivo, i messaggi di posta elettronica, gli incidenti stradali, i blackout della rete, gli attacchi di virus, gli errori nel software, le inondazioni e i terremoti sono esempi di eventi rari in cui è possibile utilizzare la distribuzione di Poisson.

Dalla (8.18) si ricava:

$$\frac{p_X(x)}{p_X(x-1)} = \frac{\lambda}{x} \qquad (x = 1, 2, \ldots), \tag{8.19}$$

così che le probabilità di Poisson (8.18) sono calcolabili in modo ricorsivo:

$$p_X(0) = e^{-\lambda}, \qquad p_X(x) = \frac{\lambda}{x}\, p_X(x-1) \qquad (x = 1, 2, \ldots).$$

Per una variabile aleatoria di Poisson si ha:

$$E(X) = \lambda, \qquad \mathrm{Var}(X) = \lambda. \tag{8.20}$$

Per il calcolo delle probabilità di Poisson in R si utilizza la funzione:

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

dove

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

- lambda vettore dei valori medi (non negativi).

Ad esempio, se $\lambda = 3$ le probabilità di Poisson per $x = 0, 1, \ldots, 10$ 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 8.11. 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),
+xlab="x",ylab="P(X=x)",type="h",main="lambda=2.5")
>
>x<-0:10
>plot(x,dpois(x,lambda=3),
+xlab="x",ylab="P(X=x)",type="h",main="lambda=3")
>
>x<-0:15
>plot(x,dpois(x,lambda=6),
+xlab="x",ylab="P(X=x)",type="h",main="lambda=6")
```

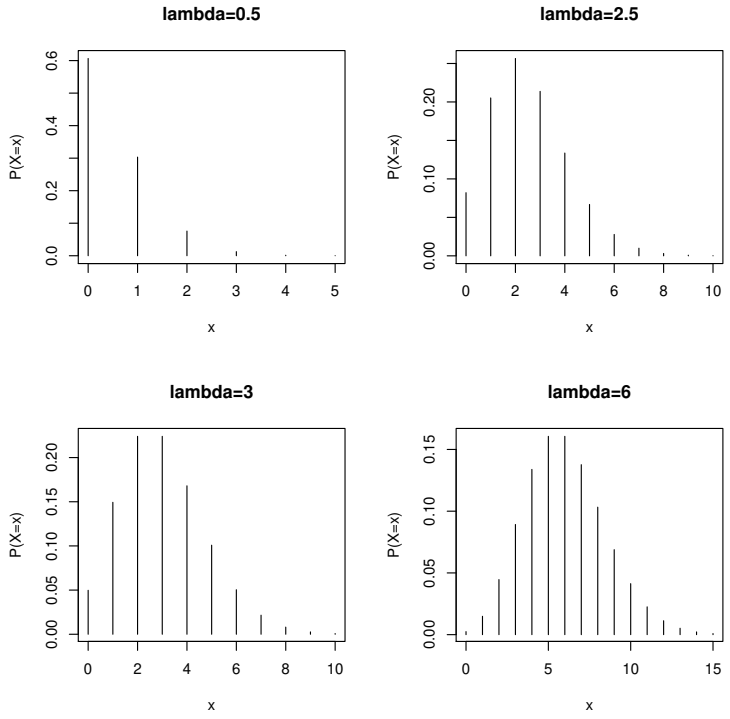

Figura 8.11: Funzione di probabilità di Poisson per alcuni valori di λ.

Per il calcolo della funzione di distribuzione di Poisson in R si utilizza la funzione:

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

dove

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

- 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 $P(X > x)$.

Ad esempio, se $\lambda = 0.5$ la funzione di distribuzione di Poisson per $x = 0, 1, \ldots, 8$ 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 8.12.

```
>par(mfrow=c(2,2))
>x<-0:5
>plot(x,ppois(x,lambda=0.5),
+xlab="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),
+xlab="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),
+xlab="x",ylab=expression(P(X<=x)),ylim=c(0,1),type="s",
+main="lambda=3")
>
>x<-0:15
>plot(x,ppois(x,lambda=6),
+xlab="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 (8.20). Ad esempio, se $\lambda = 3$ si ha $E(X) = 3$, $\mathrm{Var}(X) = 3$, $\sqrt{\mathrm{Var}(X)} = \sqrt{3}$ e $\mathrm{CV}(X) = 1/\sqrt{3}$.

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

```
qpois(z, lambda)
```

dove

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

- lambda è il vettore dei valori medi (non negativi).

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

$$P(X \leq k) \geq z \qquad (k = 0, 1, \ldots). \tag{8.21}$$

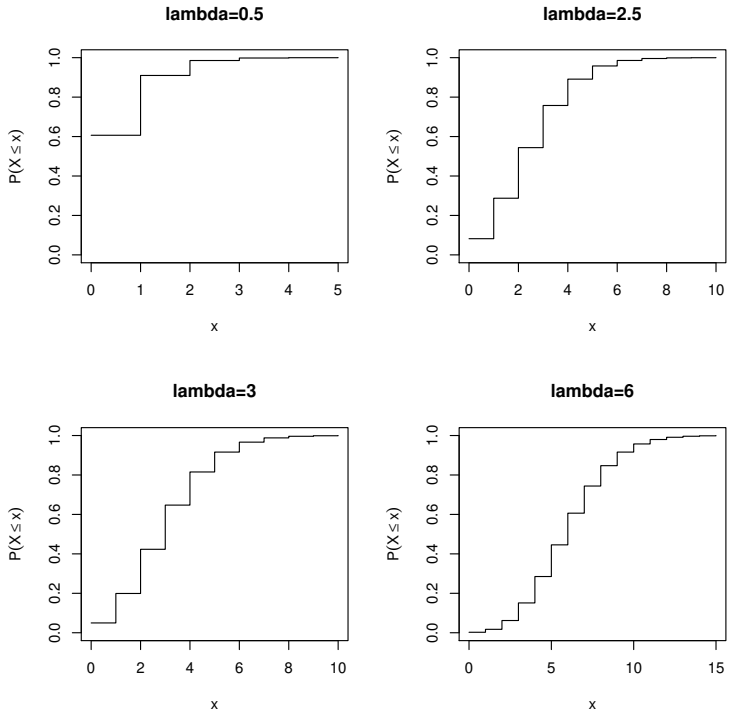

Figura 8.12: Funzione di distribuzione di Poisson per alcuni valori di λ.

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[3] 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 $0, 1, \ldots,$ 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),
```

[3] J. H. Ahrens, U. Dieter. Computer generation of Poisson deviates from modified normal distributions. ACM Transactions on Mathematical Software, 8, 163–179 (1982)

```
+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")
```

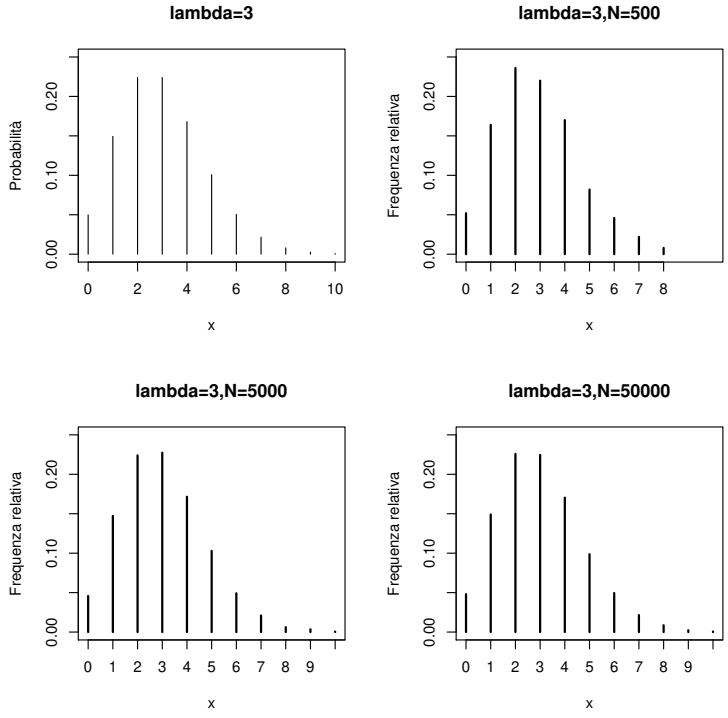

Figura 8.13: Confronto della funzione di probabilità di Poisson teorica e delle frequenze relative simulate per una variabile aleatoria di Poisson $X \sim \mathcal{P}(3)$.

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.

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), ...

Proposizione 8.1 *Sia* $X_1, X_2, \ldots$ *una successione di variabili aleatorie con* $X_n \sim \mathcal{B}(n,p)$. *Se al divergere di* n, p *tende a zero in modo tale che* $n\,p \to \lambda$, *con* $\lambda > 0$, *allora*

$$\lim_{\substack{n\to+\infty,\ p\to 0\\ n\,p\to\lambda}} p_{X_n}(k) = \lim_{\substack{n\to+\infty,\ p\to 0\\ n\,p\to\lambda}} \binom{n}{k} p^k\,(1-p)^{n-k} = \frac{\lambda^k}{k!}\,e^{-\lambda} \qquad (k = 0, 1, \ldots). \tag{8.22}$$

Un'immediata conseguenza della Proposizione 8.1 è che se il numero n di prove ripetute indipendenti di Bernoulli è elevato e se la probabilità di successo p in ogni prova è piccola, allora è possibile utilizzare la seguente approssimazione:

$$\binom{n}{k} p^k\,(1-p)^{n-k} \simeq \frac{(n\,p)^k}{k!}\,e^{-n\,p} \qquad (k = 0, 1, \ldots). \tag{8.23}$$

È opportuno menzionare che la distribuzione di Poisson costituisce una buona approssimazione della distribuzione binomiale quando nella (8.23) si ha $n \geq 20$ e $p \leq 0.05$; per $n \geq 100$ e $n\,p \leq 10$ l'approssimazione diviene poi eccellente.

Il codice seguente permette di confrontare le probabilità presenti al primo membro ed al secondo membro della (8.23) per $\lambda = 2$ e varie scelte di n e p tali che $n\,p = 2$.

```
>par(mfrow=c(2,2))
>x<-0:6
>plot(x,dbinom(x,size=10,prob=0.2),
+xlab="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),
+xlab="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),
+xlab="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),
+xlab="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 8.14 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 $n\,p = \lambda$ sia costante. Il codice seguente permette di visualizzare

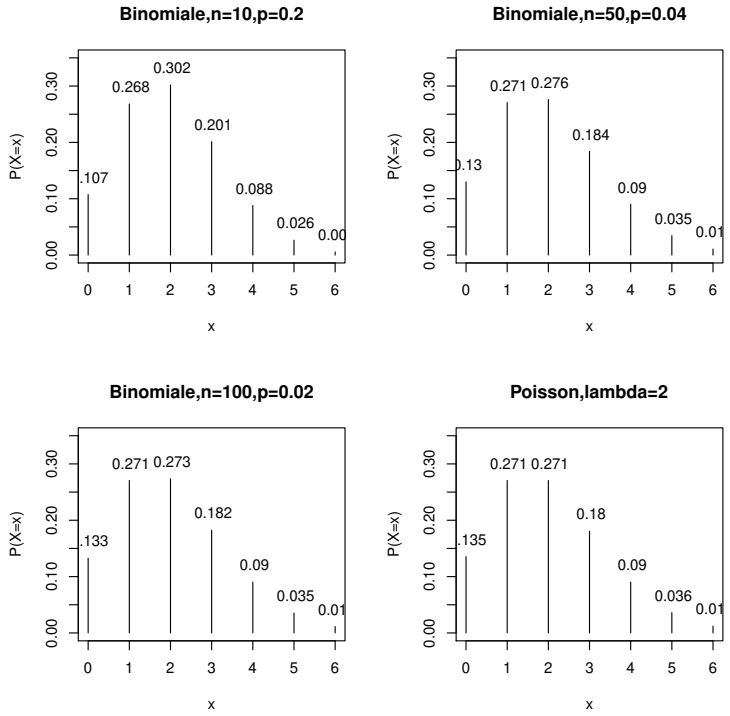

Figura 8.14: Confronto della funzione di probabilità binomiale per varie scelte di n e p tali che $n\,p = 2$ con la funzione di probabilità di Poisson con $\lambda = 2$.

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")
>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 8.15.

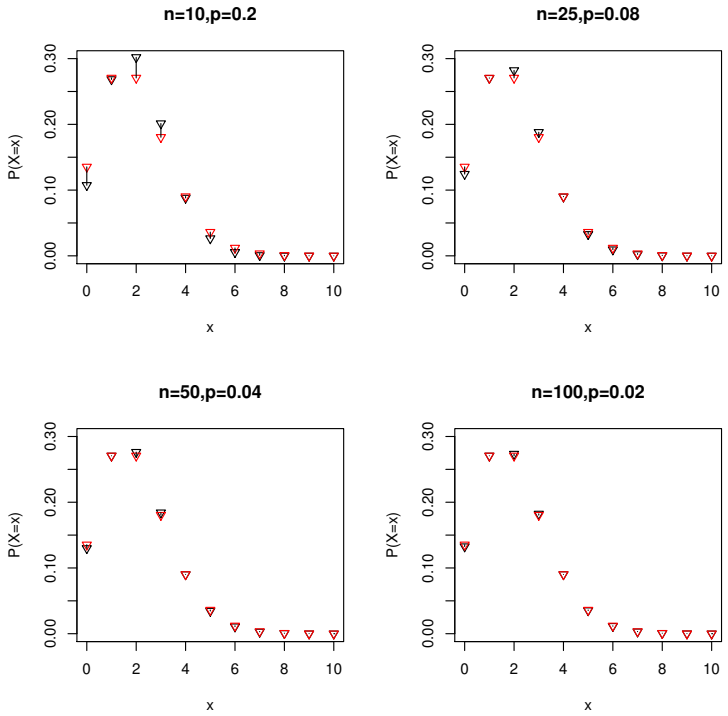

Figura 8.15: Differenze tra la funzione di probabilità binomiale per n e p tali che $np = 2$ e la funzione di probabilità di Poisson con $\lambda = 2$.

È evidente dalla Figura 8.15 che l'approssimazione della distribuzione binomiale con quella di Poisson migliora al crescere di n e al decrescere di p.

Esempio 8.4 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.

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

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

Nel caso in esame, essendo $n = 1000$ e $n\,p = 1000 \cdot 0.0004 = 0.4$, 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
```

Una proprietà molto importante di Poisson è la seguente. Se $X_1, X_2, \ldots, X_n$ sono variabili aleatorie indipendenti distribuite secondo Poisson con rispettivi parametri $\lambda_1, \lambda_2, \ldots, \lambda_n$, allora $Y_n = Y_1 + Y_2 + \ldots + Y_n$ è ancora distribuita secondo Poisson con parametro $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n$.

Esempio 8.5 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, per descrivere il numero di errori di battitura 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

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

Ciò è equivalente a richiedere che

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

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

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

Il codice seguente

```
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 8.16. Ricordiamo che ceiling(x) fornisce il più piccolo intero maggiore o uguale ad x.

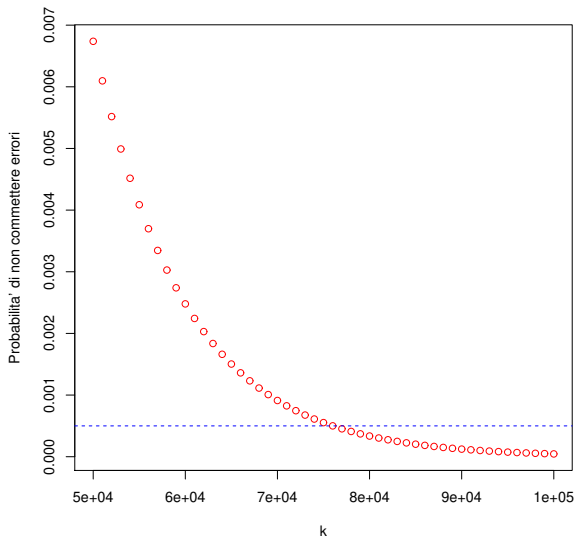

Figura 8.16: Probabilità di non commettere errori in k minuti.

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.

Se la popolazione in considerazione è descrivibile mediante una variabile aleatoria di Poisson $X \sim \mathcal{P}(\lambda)$, nei prossimi capitoli affronteremo i problemi di stimare il valore medio $E(X) = \lambda$ e di effettuare opportuni test di verifica di ipotesi sul parametro λ utilizzando un campione casuale estratto dalla stessa popolazione.

8.9 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

$$E_r = \{r \textit{ delle } k \textit{ biglie estratte sono bianche}\} \qquad (r = 0, 1, \ldots, k).$$

Facendo ricorso alla definizione classica di probabilità, si ha:

$$P(E_r) = \frac{\binom{m}{r}\binom{n}{k-r}}{\binom{m+n}{k}}, \tag{8.24}$$

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 8.8 *Una variabile aleatoria X di funzione di probabilità*

$$p_X(x) = \begin{cases} \dfrac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}, & \max\{0, k-n\} \leq x \leq \min\{m, k\} \\ 0, & \textit{altrimenti,} \end{cases} \tag{8.25}$$

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:

$$E(X) = k\,\frac{m}{m+n}, \qquad \mathrm{Var}(X) = k\frac{m\,n}{(m+n)^2}\,\frac{m+n-k}{m+n-1}\,. \tag{8.26}$$

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.

Per il calcolo delle probabilità ipergeometriche si utilizza 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 8.17.

```
>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 $x = 0, 1, \ldots, 5$. 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 $x = 0, 1, \ldots, 5$. 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 $x = 0, 1, \ldots, 20$. 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 $x = 0, 1, \ldots, 20$.

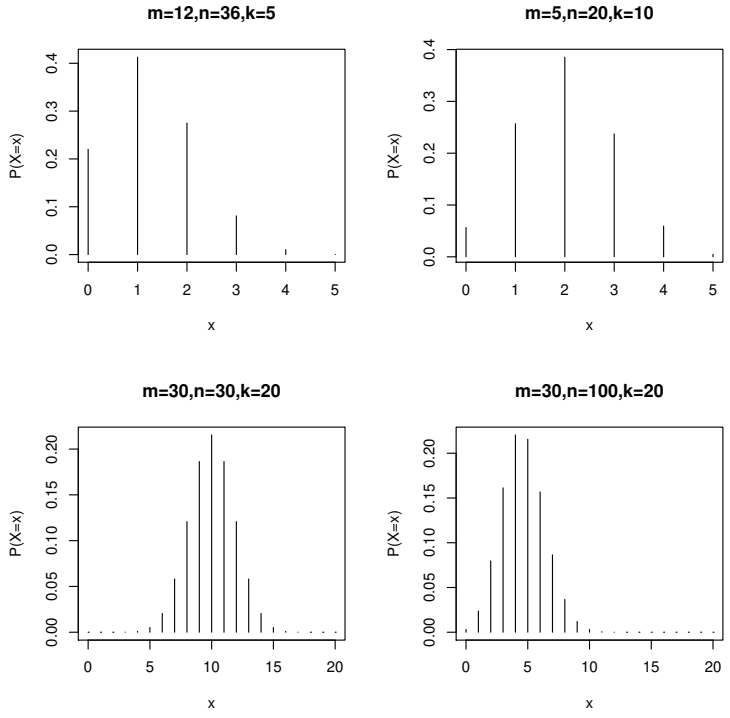

Figura 8.17: Funzione di probabilità ipergeometrica per alcuni valori di m, n e k.

Per il calcolo della funzione di distribuzione ipergeometrica si utilizza 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;

- 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:

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

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:

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

Le seguenti linee di codice permettono di visualizzare le funzioni di distrubuzione ipergeometrica di Figura 8.18.

```
>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")
```

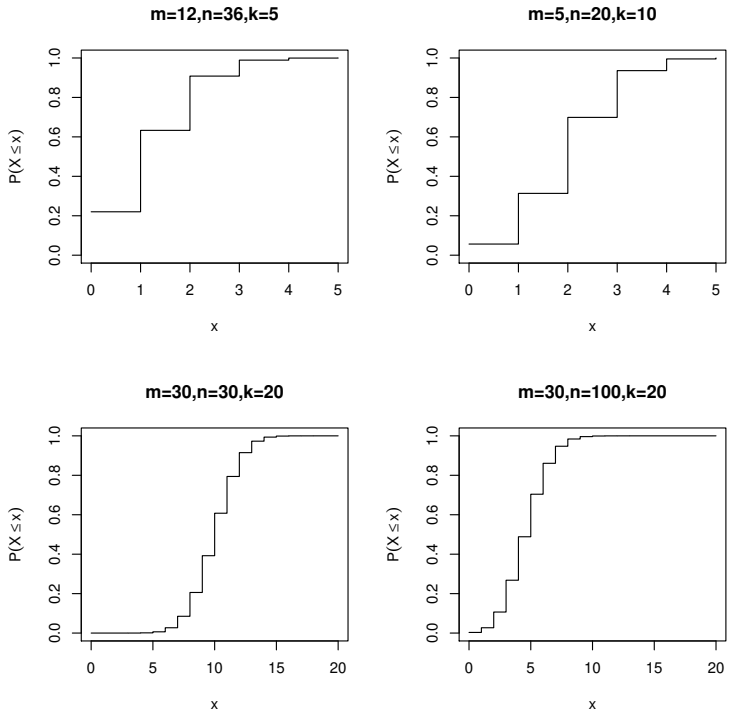

Figura 8.18: Funzione di distribuzione ipergeometrica per alcuni valori di m, n e k.

È possibile calcolare il valore medio, la varianza, la deviazione standard e il coefficiente di variazione della distribuzione ipergeometrica attraverso la (8.26). 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)
```

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;

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

$$P(X \leq r) \geq z \qquad \max\{0, k-n\} \leq r \leq \min\{m, k\}. \tag{8.27}$$

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[4] 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
```

[4] V. Kachitvichyanukul, B. Schmeiser Computer generation of hypergeometric random variates. Journal of Statistical Computation and Simulation, 22, 127–145 (1985)

dove table(sim)/length(sim) fornisce le frequenze relative con cui i numeri $0, 1, \ldots,$ 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.

Esempio 8.6 Un produttore di computer decide di acquistare monitor da una nuova start-up che rivendica severi standard di controllo della qualità. Il produttore ordina 150 monitor e decide di accettare il lotto a condizione che un campione casuale di dimensione 25 non riveli monitor difettosi. Se il lotto di 150 monitor contiene tre monitor difettosi, si desidera determinare la probabilità che il lotto venga accettato.

Denotiamo con X la variabile aleatoria che rappresenta il numero di monitor non difettosi nel campione. Si nota che $X \sim \mathcal{I}(147, 3, 25)$. La probabilità richiesta è

$$P(X = 25) = \frac{\binom{147}{25}\binom{3}{0}}{\binom{150}{25}} = 0.5764.$$

Utilizzando R otteniamo:

```
> dhyper(25,147,3,25)
[1] 0.576365
```

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

Proposizione 8.2 *Sia X una variabile aleatoria ipergeometrica descrivente l'estrazione di k biglie senza reinserimento da un'urna contenente $m + n$ biglie, di*

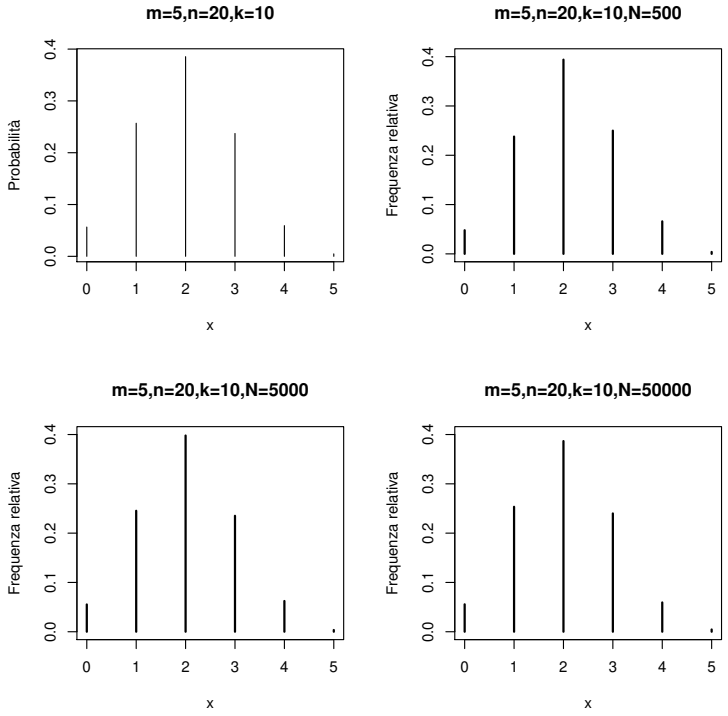

Figura 8.19: Confronto della funzione di probabilità ipergeometrica teorica e delle frequenze relative simulate per una variabile aleatoria ipergeometrica con $m = 5$, $n = 20$ e $k = 10$.

cui m sono bianche e n sono nere ($0 \leq k \leq m+n$). Se m e $m+n$ divergono in maniera tale che $m/(m+n)$ converga ad un valore $p \in (0,1)$, allora

$$\lim_{\substack{m\to+\infty,\ m+n\to+\infty \\ m/(m+n)\to p}} p_X(x) = \binom{k}{x} p^x (1-p)^{k-x} \qquad (x = 0, 1, \ldots, k). \tag{8.28}$$

Con riferimento allo schema di estrazione che ha condotto alla formula (8.24), la Proposizione 8.2 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

$$\frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}} \simeq \binom{k}{x} \left(\frac{m}{m+n}\right)^x \left(1 - \frac{m}{m+n}\right)^{k-x} \qquad (x = 0, 1, \ldots, k). \tag{8.29}$$

Il codice seguente permette di confrontare le probabilità presenti al primo membro ed al secondo membro della (8.29) 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 8.20 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.

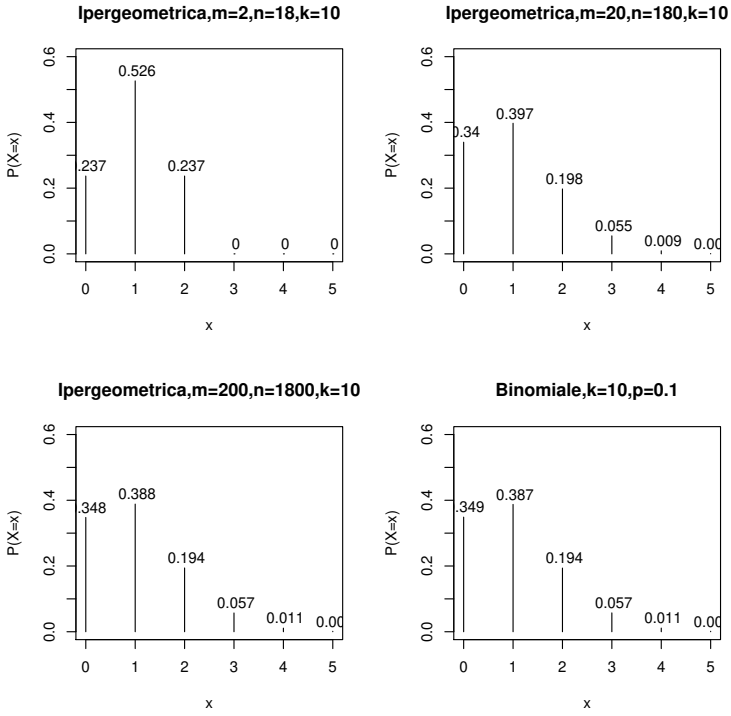

Figura 8.20: Confronto della funzione di probabilità ipergeometrica per $k = 10$ e varie scelte di m, n tali che $p = m/(n+m) = 0.1$ con la funzione di probabilità binomiale con $k = 10$ e $p = 0.1$.

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 8.21.

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

8.10 Tabelle sulle distribuzioni discrete

Concludiamo il capitolo con alcune tabelle riguardanti le variabili aleatorie discrete considerate e la loro simulazione.

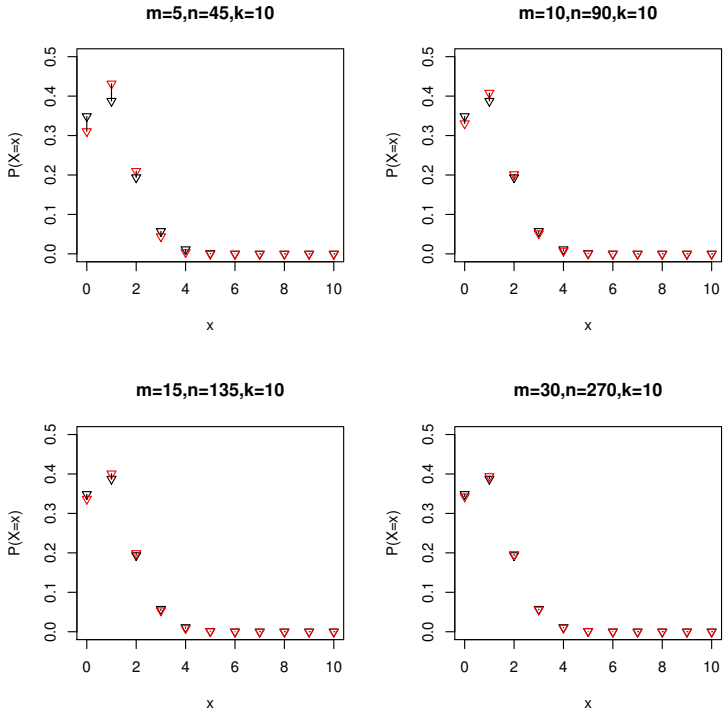

Figura 8.21: Differenze tra la funzione di probabilità ipergeometrica per $k = 10$ e varie scelte di m, n tali che $p = m/(n + m) = 0.1$ e la funzione di probabilità binomiale con $k = 10$ e $p = 0.1$.

Tabella 8.1: Funzioni di probabilità di variabili aleatorie discrete.

Distribuzione	Notazione	Funzione di probabilità
Bernoulli	$X \sim \mathcal{B}(1,p)$	$p_X(x) = \begin{cases} 1-p, & x=0 \\ p, & x=1 \\ 0, & \text{altrimenti} \end{cases} \quad (0<p<1)$
Binomiale	$X \sim \mathcal{B}(n,p)$	$p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x=0,1,\ldots,n \\ 0, & \text{altrimenti} \end{cases} \quad (n=1,2,\ldots;\ 0<p<1)$
Geometrica	$X \sim \mathcal{BN}(1,p)$	$p_X(x) = \begin{cases} p\,(1-p)^x, & x=0,1,\ldots \\ 0, & \text{altrimenti} \end{cases} \quad (0<p<1)$
Geometrica modificata	$X \sim \mathcal{BN}^*(1,p)$	$p_X(x) = \begin{cases} p\,(1-p)^{x-1}, & x=1,2,\ldots \\ 0, & \text{altrimenti} \end{cases} \quad (0<p<1)$
Binomiale negativa	$X \sim \mathcal{BN}(n,p)$	$p_X(x) = \begin{cases} \binom{n+x-1}{x} p^n (1-p)^x, & x=0,1,\ldots \\ 0, & \text{altrimenti} \end{cases} \quad (n=1,2,\ldots;\ 0<p<1)$
Binomiale negativa modificata	$X \sim \mathcal{BN}^*(n,p)$	$p_X(x) = \begin{cases} \binom{x-1}{n-1} p^n (1-p)^{x-n}, & x=n,n+1,\ldots \\ 0, & \text{altrimenti} \end{cases} \quad (n=1,2,\ldots;\ 0<p<1)$
Poisson	$X \sim \mathcal{P}(\lambda)$	$p_X(x) = \begin{cases} \dfrac{\lambda^x}{x!}\, e^{-\lambda}, & x=0,1,\ldots \\ 0, & \text{altrimenti} \end{cases} \quad (\lambda>0)$
Ipergeometrica	$X \sim \mathcal{I}(n,m,k)$	$p_X(x) = \begin{cases} \dfrac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}, & x \geq \max\{0,k-n\} \\ & x \leq \min\{m,k\} \\ 0, & \text{altrimenti} \end{cases} \quad (0 \leq k \leq m+n)$

Tabella 8.2: Valori medi, varianze e coefficienti di variazione di variabili discrete.

Nome	$E(X)$	$\mathrm{Var}(X)$	$\mathrm{CV}(X)$
Bernoulli	p	$p(1-p)$	$\sqrt{\frac{1-p}{p}}$
Binomiale	np	$np(1-p)$	$\sqrt{\frac{1-p}{np}}$
Geometrica	$(1-p)/p$	$(1-p)/p^2$	$1/\sqrt{1-p}$
Geometrica modificata	$1/p$	$(1-p)/p^2$	$\sqrt{1-p}$
Binomiale negativa	$n(1-p)/p$	$n(1-p)/p^2$	$1/\sqrt{(1-p)/n}$
Binomiale negativa modificata	n/p	$n(1-p)/p^2$	$\sqrt{(1-p)/n}$
Poisson	λ	λ	$1/\sqrt{\lambda}$
Ipergeometrica	$k\frac{m}{m+n}$	$k\frac{mn}{(m+n)^2}\frac{m+n-k}{m+n-1}$	$\sqrt{\frac{n(m+n-k)}{k\,m\,(m+n-1)}}$

Tabella 8.3: Funzioni in R per le distribuzioni discrete.

Nome	Probabilità Distribuzione	Quantili	Simulazione
Bernoulli	dbinom(x,1,prob) pbinom(x,1,prob)	qbinom(z,1,prob)	rbinom(N,1,prob)
Binomiale	dbinom(x,size,prob) pbinom(x,size,prob)	qbinom(z,size,prob)	rbinom(N,size,prob)
Geometrica	dgeom(x, prob) pgeom(x, prob)	qgeom(z, prob)	rgeom(N, prob)
Geometrica modificata	dgeom(x-1, prob) pgeom(x-1, prob)	qgeom(z, prob)+1	rgeom(N, prob)+1
Binomiale negativa	dnbinom(x, size, prob) pnbinom(x, size, prob)	qnbinom(z, size, prob)	rnbinom(x, size, prob)
Binomiale negativa modificata	dnbinom(x-n, size, prob) pnbinom(x-n, size, prob)	qnbinom(z, size, prob)+n	rnbinom(x, size, prob)+n
Poisson	dpois(x,lambda) ppois(x,lambda)	qpois(z,lambda)	rpois(N,lambda)
Ipergeometrica	dhyper(x, m, n, k) phyper(x, m, n, k)	qhyper(z, m, n, k)	rhyper(N, m, n, k)

Capitolo 9

Variabili aleatorie continue con R

9.1 Introduzione

In questo capitolo considereremo le seguenti distribuzioni continue:

- distribuzione uniforme;

- distribuzione esponenziale;

- distribuzione normale;

- distribuzione chi–quadrato;

- distribuzione di Student.

9.2 Distribuzione uniforme

Definizione 9.1 *Siano a e b numeri reali tali che $a < b$. Una variabile aleatoria X di funzione di distribuzione*

$$F_X(x) = P(X \leq x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & a \leq x < b \\ 1, & x \geq b \end{cases} \tag{9.1}$$

e corrispondente densità di probabilità

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & a < x < b \\ 0, & \textit{altrimenti} \end{cases} \tag{9.2}$$

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

La distribuzione uniforme è utilizzata in qualsiasi situazione in cui si sceglie un valore "a caso" in un fissato intervallo, senza alcuna preferenza per valori inferiori, superiori o medi nell'intervallo. Ad esempio, le posizioni degli errori in un programma, i compleanni durante un anno sono distribuiti uniformemente nei rispettivi intervalli.

La densità di probabilità uniforme traduce nel continuo il concetto di equiprobabilità nel discreto. 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:

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

Molti linguaggi di programmazione e software statistici sono dotati di un generatore di numeri pseudocasuali che produce sequenze che simulano variabili aleatorie uniformi. Una variabile aleatoria con qualsiasi altra distribuzione può essere generata a partire da una variabile aleatoria uniforme ed utilizzata poi per la simulazione al computer di vari eventi e processi.

Per il calcolo della densità uniforme in R si utilizza la funzione

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

dove

- 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 uniforme invece utilizziamo la funzione

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

dove

- 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 $P(X > x)$.

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 9.1.

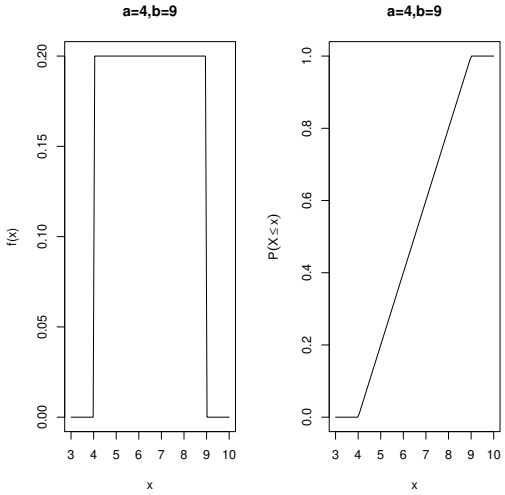

Figura 9.1: Rappresentazione della densità e della funzione di distribuzione uniforme nell'intervallo $(4, 9)$.

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

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

e corrisponde all'area del rettangolo visualizzato in Figura 9.2 visualizzata tramite il seguente codice:

```
>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
```

Per calcolare i quantili (percentili) della distribuzione uniforme nell'intervallo (a, b) in R si utilizza la funzione

```
qunif(z, min=a, max=b)
```

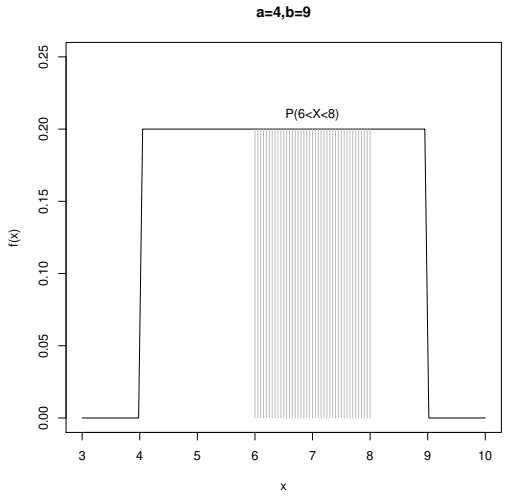

Figura 9.2: Rappresentazione della probabilità $P(6 < X < 8)$ per una variabile uniforme nell'intervallo $(4, 9)$.

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.

Il risultato della funzione è il percentile $z \cdot 100$–esimo, ossia il più piccolo numero x [illegible]nto dalla variabile aleatoria uniforme X tale che

$$F_X(x) = P(X \leq x) \geq z \qquad (a < x < b). \tag{9.3}$$

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

$$P(X \leq x) = \frac{x - 4}{5} \geq 0.25 \quad \Longleftrightarrow \quad x \geq 4 + 5 \cdot 0.25 = 5.25,$$

$$P(X \leq x) = \frac{x-4}{5} \geq 0.50 \quad \Longleftrightarrow \quad x \geq 4 + 5 \cdot 0.50 = 6.50,$$
$$P(X \leq x) = \frac{x-4}{5} \geq 0.75 \quad \Longleftrightarrow \quad x \geq 4 + 5 \cdot 0.75 = 7.75.$$

È possibile simulare la variabile aleatoria uniforme nell'intervallo (a, b) in R 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 9.1 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 9.3):

```
>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)
```

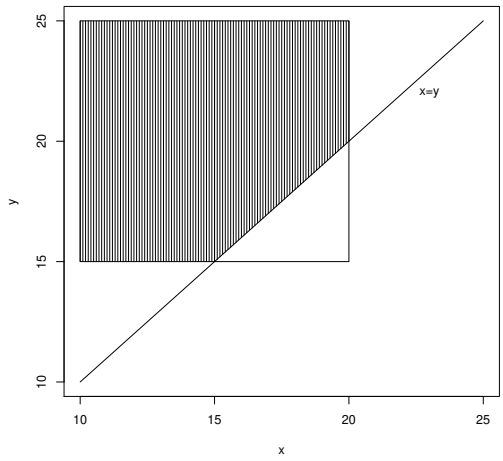

Figura 9.3: Rappresentazione del dominio $\mathcal{D}$.

Pertanto, tale probabilità può essere calcolata moltiplicando la densità congiunta per l'area tratteggiata rappresentata in Figura 9.3:

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

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 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 diff > 0 e i casi possibili sono il numero di elementi dei vettori.

Se la popolazione in considerazione è descrivibile mediante una variabile aleatoria uniforme $X \sim \mathcal{U}(0, \vartheta)$, nei prossimi capitoli affronteremo i problemi di stimare il valore medio $E(X) = \vartheta/2$ e di effettuare opportuni test di verifica di ipotesi sul parametro ϑ utilizzando un campione casuale estratto dalla stessa popolazione.

9.3 Distribuzione esponenziale

La densità di probabilità esponenziale si può interpretare 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 un tempo di attesa nel continuo.

Definizione 9.2 *Sia $\lambda > 0$. Una variabile aleatoria X di funzione di distribuzione*

$$F_X(x) = P(X \le x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \ge 0 \end{cases} \tag{9.4}$$

e corrispondente densità di probabilità

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \textit{altrimenti} \end{cases} \tag{9.5}$$

si dice esponenzialmente distribuita con parametro λ.

La variabile aleatoria 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 esponenzialmente, o allorché si considera la durata di funzionamento di componenti elettronici o di dispositivi di varia natura che si guastano per cause accidentali (corti circuiti, fulmini, scariche elettriche, imprevedibili sollecitazioni meccaniche, ...).

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

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

Il significato del parametro λ della distribuzione esponenziale può essere chiarito osservando che $E(X) = 1/\lambda$. Se X descrive un tempo, misurato in minuti, allora λ è una frequenza, misurata in 1/min. Ad esempio, se gli arrivi avvengono in media ogni mezzo minuto, allora $E(X) = 0.5$ min e $\lambda = 2$, ossia gli arrivi si verificano con una frequenza (tasso di arrivo) di 2 arrivi al minuto. Il parametro λ ha lo stesso significato del parametro λ della distribuzione di Poisson, che descrive invece il numero di arrivi.

La distribuzione esponenziale, così come la distribuzione geometrica, gode della proprietà di "assenza di memoria". Infatti, per ogni s, t reali positivi risulta:

$$P(X > s + t \,|\, X > s) = P(X > t). \tag{9.6}$$

Se si interpreta X come un tempo di attesa, la (9.6) 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.

Quando il numero di eventi (ad esempio, numero di arrivi ad un centralino telefonico, numero di arrivi ad un sistema di servizio, ...) è descritto da una distribuzione di Poisson, il tempo tra successivi eventi (ad esempio, tempi tra successivi arrivi, ...) è distribuito esponenzialmente.

Per calcolare la densità esponenziale in R si utilizza la funzione:

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

dove

- 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 esponenziale invece utilizziamo la funzione:

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

dove

- 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 9.4.
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 9.5 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)")
```

Come si evince dal grafico in Figura 9.5 la probabilità $P(0.5 < X < 1.5)$ può essere così valutata in R:

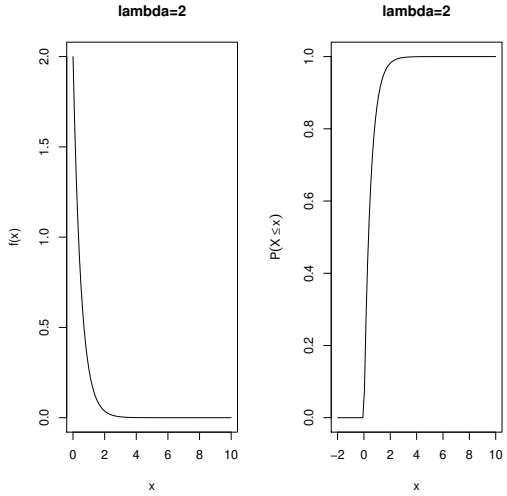

Figura 9.4: Rappresentazione della densità e della funzione di distribuzione esponenziale di valore medio 1/2.

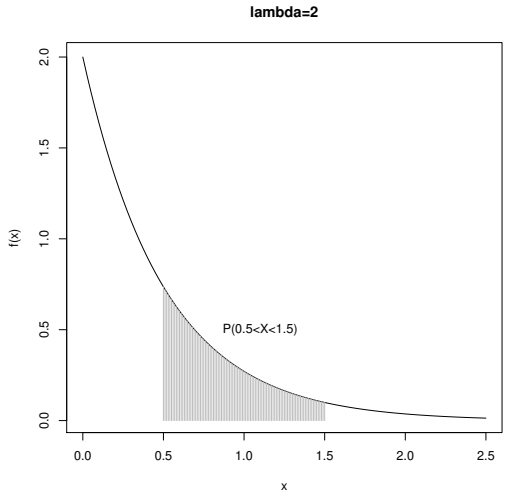

Figura 9.5: Rappresentazione della $P(0.5 < X < 1.5)$ per una variabile aleatoria esponenziale di valore medio 1/2.

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

I quantili (percentili) della distribuzione esponenziale in R si calcolano attraverso la funzione

```
qexp(z, rate = lambda)
```

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.

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 (9.3), 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

$$1 - e^{-\lambda x} \geq z \quad \Longleftrightarrow \quad x \geq -\frac{\log(1-z)}{\lambda}$$

$$\begin{aligned}
P(X \leq x) = 1 - e^{-2x} \geq 0.25 \quad &\Longleftrightarrow \quad x \geq -\frac{\log(1-0.25)}{\lambda} = 0.1438410, \\
P(X \leq x) = 1 - e^{-2x} \geq 0.50 \quad &\Longleftrightarrow \quad x \geq -\frac{\log(1-0.50)}{\lambda} = 0.3465736, \\
P(X \leq x) = 1 - e^{-2x} \geq 0.75 \quad &\Longleftrightarrow \quad x \geq -\frac{\log(1-0.75)}{\lambda} = 0.6931472.
\end{aligned}$$

È possibile simulare in R la variabile aleatoria esponenziale[1] generando una sequenza di numeri pseudocasuali mediante la funzione

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

dove

- `N` è lunghezza della sequenza da generare;

[1] J. H. Ahrens, U. Dieter. Computer methods for sampling from the exponential and normal distributions. Communications of the ACM, 15, 873–882 (1972)

- *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="Densita' 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="Densita' 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="Densita' 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="Densita' simulata,N=50000")
```

permette di confrontare in Figura 9.6 la densità esponenziale teorica di valore medio $1/2$ con la densità simulata scegliendo $N = 500, 5000, 50000$. Si nota che

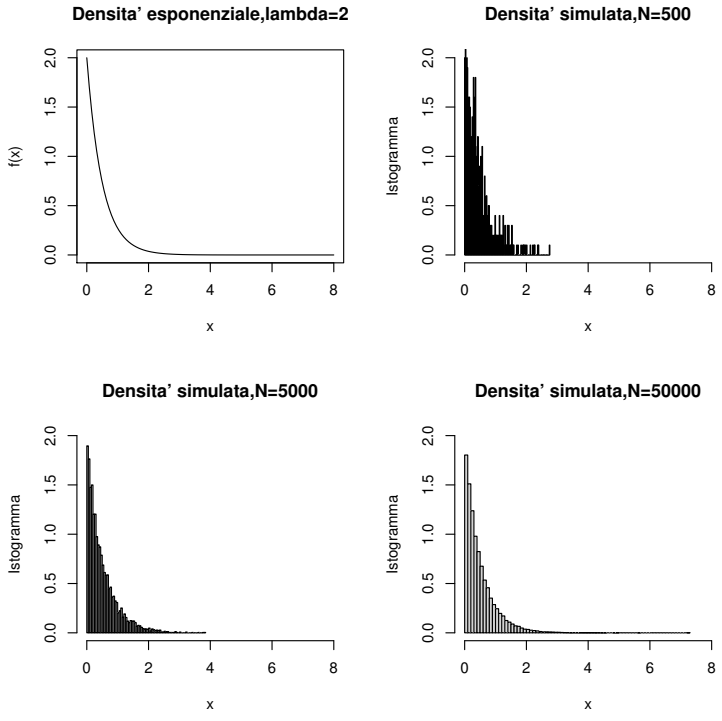

Figura 9.6: Confronto della densità esponenziale di valore medio $1/\lambda = 1/2$ con la densità simulata.

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

Se la popolazione in considerazione è descrivibile mediante una variabile aleatoria esponenziale $X \sim \mathcal{E}(\lambda)$, nei prossimi capitoli affronteremo i problemi di stimare il valore medio $E(X) = 1/\lambda$ e di effettuare opportuni test di verifica di ipotesi sul parametro λ utilizzando un campione casuale estratto dalla stessa popolazione.

9.4 Distribuzione normale

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

La distribuzione normale è spesso considerata un buon modello per variabili fisiche come peso, altezza, temperatura, voltaggio, livello di inquinamento, per analizzare gli errori di misurazione ed anche per descrivere il reddito familiare o voti degli studenti.

Definizione 9.3 *Una variabile aleatoria X di densità di probabilità*

$$f_X(x) = \frac{1}{\sigma\sqrt{2\,\pi}}\,\exp\Big\{-\frac{(x-\mu)^2}{2\,\sigma^2}\Big\}, \qquad x \in \mathbb{R} \qquad (\mu \in \mathbb{R}, \sigma > 0), \qquad (9.7)$$

si dice avere distribuzione normale di parametri μ e σ.

Dalla (9.7) 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 $\big(\sigma\sqrt{2\,\pi}\big)^{-1}$ 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*.

Per una variabile aleatoria normale il valore medio e la varianza sono:

$$E(X) = \mu, \qquad \mathrm{Var}(X) = \sigma^2.$$

Quindi, il parametro μ corrisponde al valore medio e il parametro σ alla deviazione standard.

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)
```

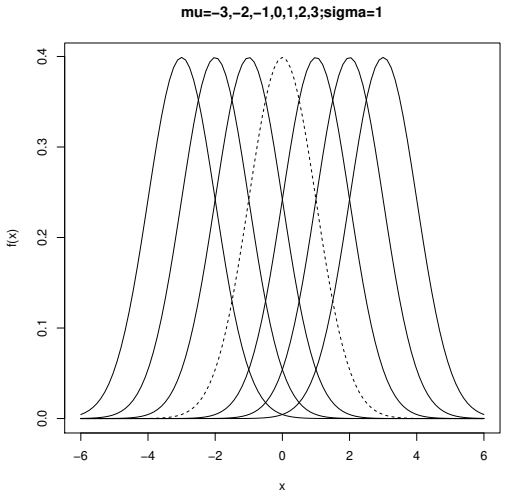

Figura 9.7: Densità normale al variare di $\mu = -3, -2, -1, 0, 1, 2, 3$ (da sinistra verso destra).

Come illustrato nella Figura 9.7 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.

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 9.8. 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.

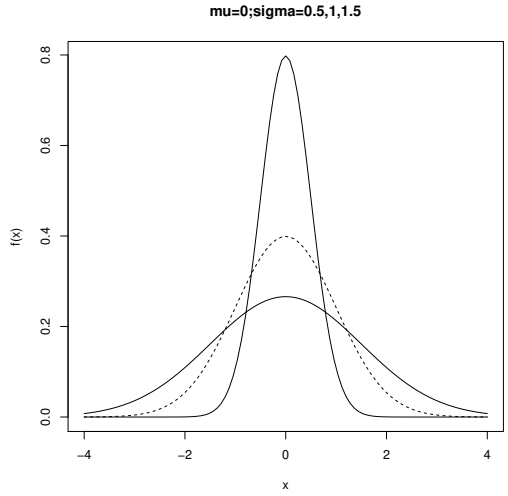

Figura 9.8: Densità normale al variare di $\sigma = 0.5, 1, 1.5$ (dall'alto verso il basso in prossimità dell'origine).

Una *variabile aleatoria normale standard*, solitamente denotata con Z, può essere ottenuta da una variabile aleatoria normale non standard $X \sim \mathcal{N}(\mu, \sigma)$ standardizzando, ossia sottraendo il valore medio e dividendo per la deviazione standard:

$$Z = \frac{X - \mu}{\sigma},$$

da cui segue

$$X = \mu + \sigma\, Z$$

Usando questa trasformazione, qualsiasi variabile aleatoria normale può essere ottenuta da una variabile aleatoria normale standard Z.

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

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(y)\, dy = \Phi\Big(\frac{x - \mu}{\sigma}\Big), \qquad x \in \mathbb{R} \tag{9.8}$$

dove

$$\Phi(z) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{z} \exp\Big\{-\frac{y^2}{2}\Big\}\, dy\,, \qquad z \in \mathbb{R}. \tag{9.9}$$

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

$$P(a < X < b) = F_X(b) - F_X(a) = \Phi\Big(\frac{b-\mu}{\sigma}\Big) - \Phi\Big(\frac{a-\mu}{\sigma}\Big). \tag{9.10}$$

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 9.9. La funzione arrows() ha come argomenti 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.

$\Rightarrow$ Regola del $3\,\sigma$

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

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

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

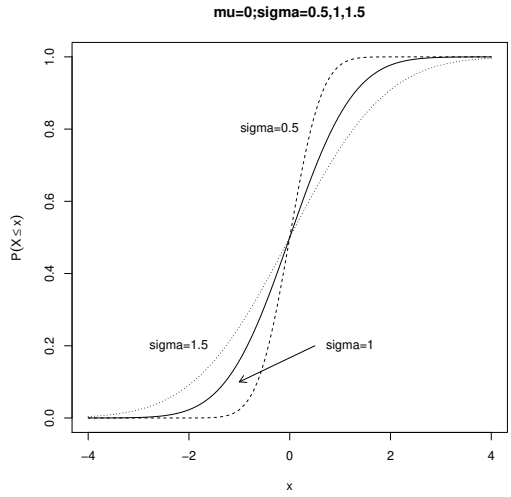

Figura 9.9: Funzione di distribuzione normale al variare di $\sigma = 0.5, 1, 1.5$.

La regola del $3\,\sigma$ 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)
```

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.

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 (9.3), ossia che $F_X(x) = P(X \leq x) \geq z$. 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 9.2 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

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

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 $\varepsilon = 1.959964 \cdot \sigma$. ◊

Esempio 9.3 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 $p = 0.9544997$. ◊

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="Densita' 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="Densita' 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="Densita' 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="Densita' simulata,N=50000")
```

permette di confrontare in Figura 9.10 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.

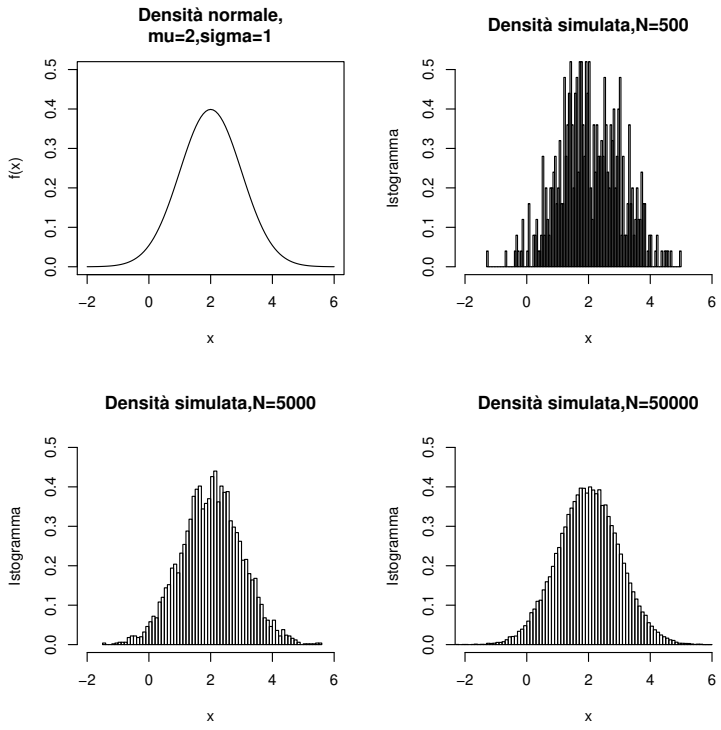

Figura 9.10: Confronto della densità normale con $\mu = 2$ e $\sigma = 1$ con la densità simulata.

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

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

ha distribuzione normale con valore medio e varianza:

$$E(Y) = a_1 \mu_1 + a_2 \mu_2 + \ldots + a_n \mu_n$$
$$\text{Var}(Y) = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \ldots + a_n^2 \sigma_n^2.$$

In particolare, se $X_1, X_2, \ldots, X_n$ sono variabili aleatorie normali indipendenti con $X_i \sim \mathcal{N}(\mu, \sigma)$ per $i = 1, 2, \ldots, n$ e se $a_1 = a_2 = \ldots = a_n = 1/n$, allora Y

coincide con la media campionaria

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

che ha una distribuzione normale con valore medio e varianza:

$$E(\overline{X}_n) = \mu, \qquad \mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}\cdot$$

Questo risultato mostra che se la popolazione è descritta da una variabile aleatoria normale $X \sim \mathcal{N}(\mu, \sigma)$ e da essa si estrae un campione $X_1, X_2, \ldots, X_n$, allora la media campionaria $\overline{X}_n$ è anche normale con valore medio μ e varianza σ^2/n. Inoltre, se la popolazione è descritta da una variabile aleatoria normale $X \sim \mathcal{N}(\mu, \sigma)$, la variabile aleatoria standardizzata

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

è normale standard.

Se la popolazione in considerazione è descrivibile mediante una variabile aleatoria normale $X \sim \mathcal{N}(\mu, \sigma)$, nei prossimi capitoli affronteremo i problemi di stimare il valore medio $E(X) = \mu$ e la varianza $\mathrm{Var}(X) = \sigma^2$ e di effettuare opportuni test di verifica di ipotesi sui parametri μ e σ utilizzando un campione casuale estratto dalla stessa popolazione.

⇒ Approssimazione della distribuzione binomiale con la distribuzione normale

Il calcolo delle probabilità binomiali diviene rapidamente oneroso al crescere del numero n delle prove. È quindi utile ricercare delle formule approssimate in grado di rendere agevole tale calcolo e, al contempo, accettabile l'errore derivante dall'approssimazione.

Abraham de Moivre (1667–1754) ottenne la prima versione del teorema centrale del limite come approssimazione della distribuzione binomiale.

Teorema 9.1 (Teorema di De Moivre-Laplace) *Sia $X_1, X_2, \ldots$ una successione di variabili aleatorie indipendenti distribuite alla Bernoulli con parametro p $(0 < p < 1)$, e sia $Y_n = X_1 + X_2 + \ldots + X_n$. Allora per ogni $x \in \mathbb{R}$ risulta:*

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

ossia

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

converge in distribuzione alla variabile aleatoria Z normale standard.

Ricordiamo che se $X_1, X_2, \ldots$ sono variabili aleatorie indipendenti di Bernoulli di parametro p, allora

$$Y_n = X_1 + X_2 + \ldots + X_n$$

è una *variabile aleatoria binomiale di valore medio $n\,p$ e varianza $n\,p\,(1-p)$*. Il Teorema 9.1 mostra che sottraendo a Y_n la sua media $n\,p$ e dividendo la differenza per la deviazione standard $\sqrt{n\,p\,(1-p)}$, 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

$$Y_n \simeq n\,p + \sqrt{n\,p\,(1-p)}\;Z, \tag{9.11}$$

al variare di n con p fissato. Si noti che il secondo membro della (9.11), ossia $n\,p + \sqrt{n\,p\,(1-p)}\;Z$, è una variabile aleatoria con densità normale di valore medio $n\,p$ e varianza $n\,p\,(1-p)$. Il codice seguente confronta la densità normale di valore medio $n\,p$ e varianza $n\,p\,(1-p)$ 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,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=25,p=0.2")
>lines(x,dbinom(x,n,0.2),type="h")
>
>x<-0:50
>n<-50
>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=50,p=0.2")
>lines(x,dbinom(x,n,0.2),type="h")
>
>x<-0:75
>n<-75
>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=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 9.11. Si nota che l'approssimazione migliora al crescere di n.

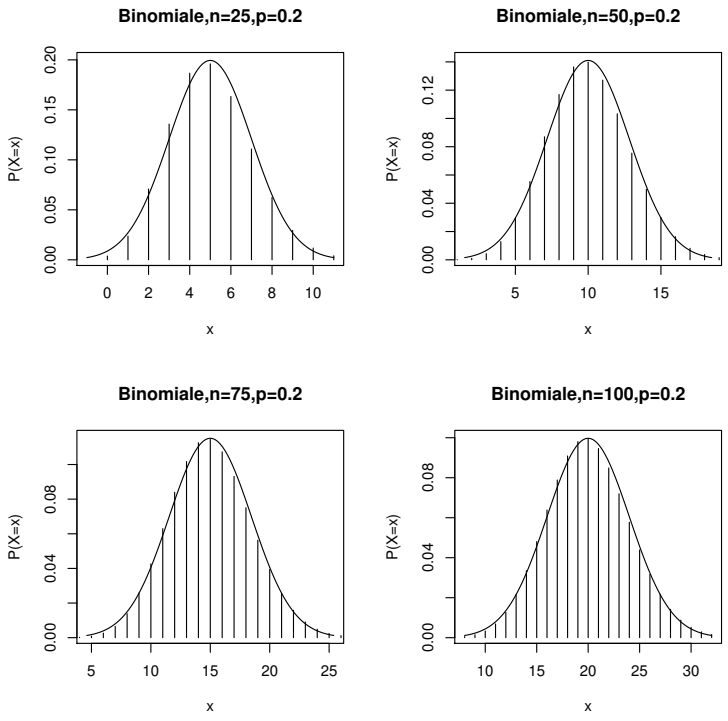

Figura 9.11: Confronto della probabilità binomiale della variabile $Y_n \sim \mathcal{B}(n, 0.2)$ con la densità normale di valor medio $\mu = n\,p$ e deviazione standard $\sigma = \sqrt{np(1-p)}$ per varie scelte di n.

Esaminiamo ora l'approssimazione (9.11) della binomiale alla normale al variare di p con n fissato. Il codice seguente confronta la densità normale di valore medio $n\,p$ e varianza $n\,p\,(1-p)$ 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 9.12. 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 $p = 1/2$.

⇒ Approssimazione della distribuzione di somme di variabili aleatorie indipendenti

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 della distribuzione della somma di variabili aleatorie indipendenti, evidenziando al contempo la grande importanza della distribuzione normale nella statistica inferenziale.

Teorema 9.2 (Teorema centrale di convergenza) *Sia $X_1, X_2, \ldots$ 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 $Y_n = X_1 + X_2 + \ldots + X_n$, per*

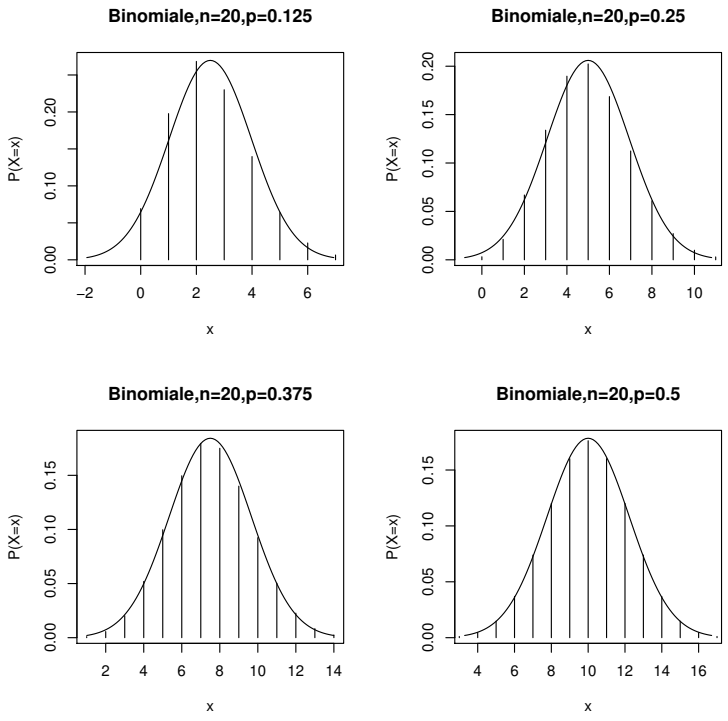

Figura 9.12: Confronto della probabilità binomiale della variabile $Y_n \sim \mathcal{B}(20, p)$ con la densità normale di valor medio $\mu = n\,p$ e deviazione standard $\sigma = \sqrt{np(1-p)}$ per varie scelte di p.

ogni $x \in \mathbb{R}$ *risulta:*

$$\lim_{n\to+\infty} P\left(\frac{Y_n - n\,\mu}{\sigma\,\sqrt{n}} \le x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\,dy = \Phi(x), \tag{9.12}$$

ossia la successione delle variabili aleatorie standardizzate

$$\frac{Y_n - E(Y_n)}{\sqrt{\mathrm{Var}(Y_n)}} = \frac{Y_n - n\,\mu}{\sigma\,\sqrt{n}} \xrightarrow{d} Z, \tag{9.13}$$

converge in distribuzione alla variabile aleatoria normale standard.

Il Teorema 9.2 mostra inoltre che sottraendo a $Y_n = X_1 + X_2 + \ldots + X_n$ 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 della somma

$$Y_n = X_1 + X_2 + \ldots + X_n$$

è *approssimativamente normale con valore medio* $n\,\mu$ *e varianza* $n\,\sigma^2$, ossia

$$Y_n \simeq n\,\mu + \sigma\,\sqrt{n}\,Z.$$

Se denotiamo con

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

la media campionaria, la (9.13) mostra che

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z, \tag{9.14}$$

converge in distribuzione alla variabile aleatoria normale standard. Quindi, per n grande la distribuzione della media campionaria $\overline{X}_n$ è *approssimativamente normale con valore medio* μ *e varianza* σ^2/n, ossia

$$\overline{X}_n \simeq \mu + \frac{\sigma}{\sqrt{n}}\,Z.$$

Quindi, se la popolazione da cui è estratto il campione è descritta da una variabile aleatoria discreta o continua X, con valore medio μ e varianza σ^2, entrambi finiti, allora la media campionaria $\overline{X}_n$, per campioni numerosi, è approssimativamente normale con valore medio μ e varianza σ^2/n. Questo risultato gioca un ruolo fondamentale nella statistica inferenziale e, in particolare, nella stima intervallare e nella verifica delle ipotesi.

Va menzionato che la bontà delle approssimazioni dipende da n e dal tipo di distribuzione delle variabili $X_1, X_2, \ldots, X_n$. L'approssimazione migliora al crescere di n e nelle applicazioni spesso si verifica che essa è già soddisfacente per $n \ge 30$.

⇒ Approssimazione della distribuzione di Poisson con la distribuzione normale

Se, ad esempio, supponiamo che $X_1, X_2, \ldots$ è una *successione di variabili aleatorie indipendenti di Poisson* ognuna di parametro λ, allora $Y_n = X_1 + X_2 + \ldots + X_n$ è ancora una variabile aleatoria di Poisson di parametro $n\,\lambda$, e quindi $E(Y_n) = n\,\lambda$ e $\mathrm{Var}(Y_n) = n\,\lambda$. Il teorema centrale di convergenza afferma che per n grande la distribuzione di $Y_n = X_1 + X_2 + \ldots + X_n$ è approssimativamente normale con valore medio $n\,\lambda$ e varianza $n\,\lambda$, ossia

$$Y_n \simeq n\lambda + \sqrt{n\,\lambda}\,Z,$$

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 della variabile Y_n parametro $n\,\lambda$ alla normale di valore medio $n\,\lambda$ 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),
+xlab="x",ylab="P(X=x)",main="Poisson, n lambda=5")
>lines(x,dpois(x, 5),type="h")
>
>curve(dnorm(x,10,sqrt(10)),from=10-3*sqrt(10),to=10+3*sqrt(10),
+xlab="x",ylab="P(X=x)",main="Poisson, n lambda=10")
>lines(x,dpois(x, 10),type="h")
>
>curve(dnorm(x,25,sqrt(25)),from=25-3*sqrt(25),to=25+3*sqrt(25),
+xlab="x",ylab="P(X=x)",
>main="Poisson,n lambda=25")
>lines(x,dpois(x, 25),type="h")
>
>curve(dnorm(x,50,sqrt(50)),from=50-3*sqrt(50),to=50+3*sqrt(50),
+xlab="x",ylab="P(X=x)",main="Poisson, n lambda=50")
>lines(x,dpois(x, 50),type="h")
```

permette di visualizzare la Figura 9.13 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.

Il teorema centrale di convergenza o teorema centrale del limite giocherà un ruolo fondamentale nella stima intervallare del valore medio per popolazioni discrete o continue (non normali) considerando campioni numerosi estratti dalla popolazione in esame.

9.5 Distribuzione chi–quadrato

La distribuzione chi–quadrato fu introdotta intorno al 1900 da un famoso matematico inglese Karl Pearson (1857-1936), considerato uno dei fondatori della statistica matematica. Pearson era un insegnante e collaboratore di William

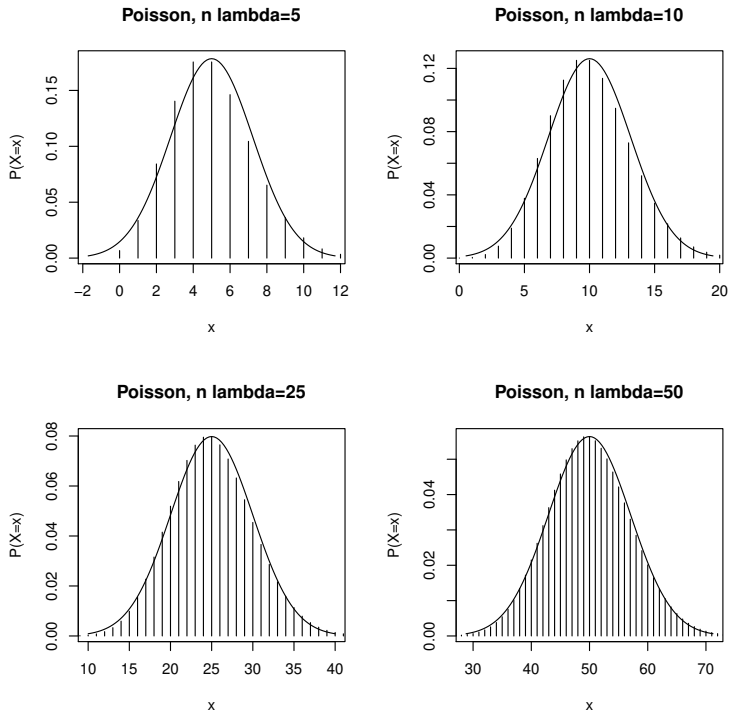

Figura 9.13: Confronto della probabilità di Poisson della variabile $X \sim \mathcal{P}(n\,\lambda)$ con la densità normale di valor medio $\mu = n\,\lambda$ e deviazione standard $\sigma = \sqrt{n\,\lambda}$ per varie scelte di $n\,\lambda$.

Gosset (che firmava i suoi lavori con lo pseudonimo Student). Tale distribuzione riveste un ruolo importante nell'inferenza statistica, in particolare nella stima intervallare della varianza di una popolazione normale, ed anche in molti test di verifica di ipotesi statistiche.

Per definire la densità chi-quadrato occorre introdurre la *funzione gamma*, così definita:

$$\Gamma(\nu) = \int_0^{+\infty} x^{\nu-1}\, e^{-x}\, dx, \qquad \nu > 0, \tag{9.15}$$

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

$$\Gamma(\nu) = (\nu - 1)\, \Gamma(\nu - 1), \qquad \nu > 1. \tag{9.16}$$

La funzione gamma è una generalizzazione dei fattoriali; infatti, se ν è un intero positivo, usando iterativamente la (9.16), si ottiene:

$$\Gamma(\nu) = (\nu - 1)!, \qquad \nu = 1, 2, \ldots,$$

avendo fatto uso della proprietà $\Gamma(1) = 1$ direttamente ricavata dalla (9.15).

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 9.4 *Una variabile aleatoria X di densità di probabilità*

$$f_X(x) = \begin{cases} \dfrac{1}{\Gamma(n/2)} \left(\dfrac{1}{2}\right)^{n/2} x^{n/2-1}\, e^{-x/2}, & x > 0 \\ 0, & x \le 0 \end{cases} \tag{9.17}$$

con n intero positivo e con $\Gamma(\nu)$ definita in (9.15), 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. Si nota che la densità (9.17) dipende soltanto dal numero n di gradi di libertà.

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 $x = n - 2$. In particolare, quando $n = 2$, (9.17) corrisponde ad una densità esponenziale di valore medio $1/\lambda = 2$.

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

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

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

Teorema 9.3 *Siano $X_1, X_2, \ldots, X_n$ variabili aleatorie indipendenti, con $X_i \sim \mathcal{N}(0,1)$ per $i = 1, 2, \ldots, n$. Allora,*

$$Y_n = X_1^2 + X_2^2 + \ldots + X_n^2$$

ha distribuzione chi–quadrato con n gradi di libertà.

Il Teorema 9.3 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.

Per il calcolo della densità chi-quadrato in R si utilizza la funzione

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

dove

- 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)
```

dove

- 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 9.14 la densità di probabilità e la funzione di distribuzione di una variabile aleatoria $X \sim \chi^2(n)$ per $n = 1, 3, 5, 7$. Si nota che la densità è decrescente per $n = 1$ e presenta un unico punto di massimo in $n - 2$ quando $n = 3, 5, 7$.

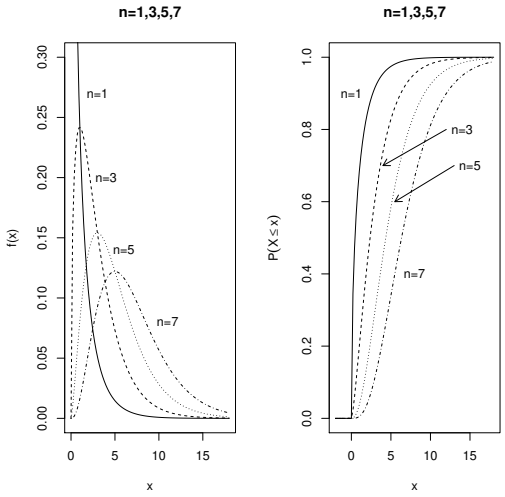

Figura 9.14: Densità di probabilità e funzione di distribuzione di $X \sim \chi^2(n)$.

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

```
qchisq(z, df)
rchisq(N, df)
```

dove df indica il numero di gradi di libertà.

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

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

ha densità chi–quadrato con $k = k_1 + k_2 + \ldots + k_n$ gradi di libertà.

Se la popolazione è descritta da una variabile aleatoria normale $X \sim \mathcal{N}(\mu, \sigma^2)$ e si considera un campione $X_1, X_2, \ldots, X_n$ estratto dalla popolazione, la variabile aleatoria

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

denota la varianza campionaria. Nella statistica inferenziale, si dimostra che per una popolazione descritta da una variabile aleatoria normale $X \sim \mathcal{N}(\mu, \sigma^2)$ la variabile aleatoria

$$Q_n = \frac{(n-1)\, S_n^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2. \tag{9.18}$$

è distribuita con legge chi-quadrato con $n-1$ gradi di libertà. Infatti,

$$Q_n = \sum_{i=1}^{n} \Big(\frac{X_i - \mu}{\sigma} - \frac{\overline{X}_n - \mu}{\sigma} \Big)^2 = \sum_{i=1}^{n} \Big(\frac{X_i - \mu}{\sigma} \Big)^2 + \sum_{i=1}^{n} \Big(\frac{\overline{X}_n - \mu}{\sigma} \Big)^2$$
$$-2\, \frac{\overline{X}_n - \mu}{\sigma} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} = \sum_{i=1}^{n} \Big(\frac{X_i - \mu}{\sigma} \Big)^2 - \Big(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \Big)^2,$$

che si presenta come la differenza tra due variabili aleatorie chi-quadrato, la prima con n gradi di libertà e la seconda con un solo grado di libertà.

9.6 Distribuzione di Student

Un'altra distribuzione di considerevole interesse applicativo è quella di Student. Student è lo pseudonimo con cui il matematico inglese William S. Gosset (1876–1937) pubblicava i suoi articoli. Lavorando per il birrificio irlandese Guinness, Gosset derivò tale distribuzione da lui utilizzata nei problemi di controllo della qualità nella produzione di birra.

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

Definizione 9.5 *Una variabile aleatoria X di densità di probabilità*

$$f_X(x) = \frac{\Gamma\Big(\frac{n+1}{2}\Big)}{\sqrt{n\,\pi}\,\Gamma\Big(\frac{n}{2}\Big)} \Big(1 + \frac{x^2}{n}\Big)^{-(n+1)/2}, \qquad x \in \mathbb{R} \tag{9.19}$$

con n intero positivo e con $\Gamma(\nu)$ definita in (9.15), 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. Si nota che la densità (9.19) dipende soltanto dal numero n di gradi di libertà.

Notiamo che la densità (9.19) è unimodale, simmetrica intorno all'asse $x = 0$ e dipendente dal solo parametro rappresentante il numero di gradi di libertà. Per $n = 1$ la (9.19) si identifica con la *densità di Cauchy*

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

Per calcolare la densità di Student in R utilizziamo la funzione:

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

dove

- 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 = Inf è consentito).

Per calcolare la funzione di distribuzione di Student utilizziamo la funzione:

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

dove

- 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 $P(X \le x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

Al limite per $n \to +\infty$ la densità di Student converge alla densità normale standard. Infatti, risulta

$$\lim_{n\to+\infty} \frac{\Gamma\Big(\frac{n+1}{2}\Big)}{\sqrt{n\,\pi}\,\Gamma\Big(\frac{n}{2}\Big)}\,\Big(1+\frac{x^2}{n}\Big)^{-(n+1)/2} = \frac{1}{\sqrt{2\pi}}\,\exp\Big\{-\frac{x^2}{2}\Big\}, \qquad x \in \mathbb{R}.$$

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")
```

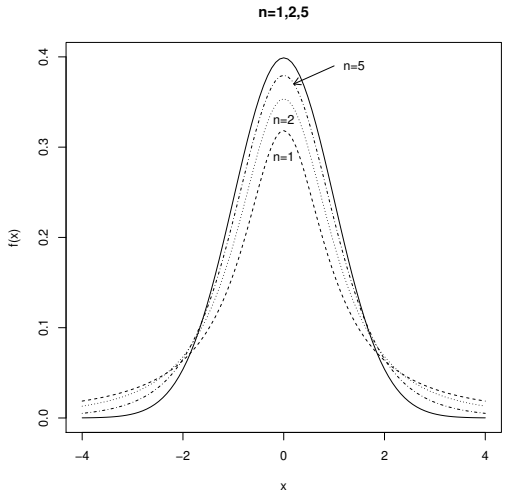

Figura 9.15: Densità di probabilità di $X \sim \mathcal{T}(n)$ per $n = 1, 2, 5$ e densità normale standard (curva con tratto continuo)

permette di rappresentare in Figura 9.15 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). La densità di Student è una curva simmetrica a campana che può essere facilmente confusa con la densità normale standard. Come si evince in Figura 9.15, il picco della densità di Student è più basso e le sue code sono più allungate rispetto alla densità normale standard. Si nota poi che al crescere di n la densità di Student tende alla densità normale standard.

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 $n = 2, 3, \ldots$. Inoltre la varianza non esiste se $n = 1$ e diverge se $n = 2$; per $n = 3, 4, \ldots$ si ha poi:

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

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

```
qt(p, df)
rt(N, df)
```

dove `df` indica il numero di gradi di libertà.

Il teorema seguente mostra la stretta connessione esistente tra una variabile a distribuzione di Student (9.19) e variabili a distribuzione chi–quadrato e normale standard.

Teorema 9.4 *Siano $Y \sim \chi^2(n)$ e $Z \sim \mathcal{N}(0,1)$ variabili aleatorie indipendenti. Allora*

$$X = \frac{Z}{\sqrt{Y/n}} \tag{9.20}$$

ha distribuzione di Student con n gradi di libertà.

Se la popolazione è descritta da una variabile aleatoria normale $X \sim \mathcal{N}(\mu, \sigma^2)$ e si considera un campione $X_1, X_2, \ldots, X_n$ estratto dalla popolazione, la variabile aleatoria

$$T_n = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}},$$

è distribuita con legge di Student con $n-1$ gradi di libertà. Infatti, ricordando la (9.18), si ha

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \frac{\sigma/\sqrt{n}}{S_n/\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \frac{1}{\sqrt{\frac{(n-1)\,S_n^2}{\sigma^2\,(n-1)}}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \frac{1}{\sqrt{\frac{Q_n}{n-1}}}$$

che è della forma (9.20). Per il Teorema 9.4, T_n è quindi distribuita con legge di Student con $n-1$ gradi di libertà.

In particolare, in questo capitolo abbiamo mostrato che se $X_1, X_2, \ldots, X_n$ sono variabili aleatorie indipendenti e normali con valore medio μ e varianza σ^2, allora

- $Z_n = \dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ è distribuita con legge normale standard;
- $T_n = \dfrac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$ è distribuita con legge di Student con $n-1$ gradi di libertà;
- $Q_n = \dfrac{(n-1)\,S_n^2}{\sigma^2}$ è distribuita con legge chi-quadrato con $n-1$ gradi di libertà.

Inoltre, se $X_1, X_2, \ldots, X_n$ sono variabili aleatorie indipendenti e identicamente distribuite con valore medio finito μ e varianza finita σ^2, allora per il teorema centrale di convergenza

- $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z$, ossia converge in distribuzione alla variabile aleatoria normale standard.

Nel seguito, utilizzeremo le distribuzioni di probabilità *normale standard, chi-quadrato e di Student* per la stima intervallare dei parametri non noti di una popolazione e nei test statistici di verifica di ipotesi sui parametri e sulle distribuzioni sia per variabili aleatorie discrete che per variabili aleatorie continue.

Concludiamo il capitolo con alcune tabelle riguardanti le variabili aleatorie continue considerate e la loro simulazione utilizzando il linguaggio R.

9.7 Tabelle per le distribuzioni continue

Tabella 9.1: Densità di probabilità di variabili aleatorie continue.

Distribuzione	Notazione	Funzione densità di probabilità
Uniforme	$X \sim \mathcal{U}(a,b)$	$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & a < x < b \\ 0, & \text{altrimenti} \end{cases}$ $(a, b \in \mathbb{R})$
Esponenziale	$X \sim \mathcal{E}(1,\lambda)$	$f_X(x) = \begin{cases} \lambda\, e^{-\lambda\, x}, & x > 0 \\ 0, & \text{altrimenti} \end{cases}$ $(\lambda > 0)$
Normale	$X \sim \mathcal{N}(\mu,\sigma)$	$f_X(x) = \dfrac{1}{\sigma\sqrt{2\,\pi}} \exp\left\{ -\dfrac{(x-\mu)^2}{2\,\sigma^2} \right\}, \quad x \in \mathbb{R}$ $(\mu \in \mathbb{R}, \sigma > 0)$
Chi-quadrato	$X \sim \chi^2(n)$	$f_X(x) = \begin{cases} \dfrac{1}{2^{n/2}\,\Gamma\left(\dfrac{n}{2}\right)}\, e^{-x/2}\, x^{n/2-1}, & x > 0 \\ 0, & \text{altrimenti} \end{cases}$ $(n = 1, 2, \ldots)$
Student	$X \sim \mathcal{T}(n)$	$f_X(x) = \dfrac{\Gamma\left(\dfrac{n+1}{2}\right)}{\sqrt{n\,\pi}\,\Gamma\left(\dfrac{n}{2}\right)} \left(1 + \dfrac{x^2}{n}\right)^{-(n+1)/2}, \quad x \in \mathbb{R}$ $(n = 1, 2, \ldots)$

Tabella 9.2: Valori medi, varianze e coefficienti di variazione di variabili continue.

Distribuzione	Valore medio $E(X)$	Varianza $\mathrm{Var}(X)$	Coefficiente di variazione $CV(X)$
Uniforme	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{b-a}{\sqrt{3}\,(a+b)}$
Esponenziale	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	1
Normale	μ	σ^2	$\frac{\sigma}{\mu}$
Chi-quadrato	n	$2\,n$	$\sqrt{\frac{2}{n}}$
Student	0 se $n > 1$	$+\infty$ se $n = 2$ $\frac{n}{n-2}$ se $n = 3, 4, \ldots$	

Tabella 9.3: Funzioni in R per le distribuzioni continue.

Nome	Densità Distribuzione Quantili	Simulazione
Uniforme	dunif(x,min=a,max=b) punif(x,min=a,max=b) qunif(z,min=a,max=b)	runif(N,min=a,max=b)
Esponenziale	dexp(x,rate=lambda) pexp(x,rate=lambda) qexp(z,rate=lambda)	rexp(N,rate=lambda)
Normale	dnorm(x,mean=mu,sd=sigma) pnorm(x,mean=mu,sd=sigma) qnorm(z,mean=mu,sd=sigma)	rnorm(N,mean=mu,sd=sigma)
Chi-quadrato	dchisq(x, df) pchisq(x, df) qchisq(z, df)	rchisq(N, df)
Student	dt(x, df) pt(x, df) qt(z, df)	rt(N, df)

Capitolo 10

Stima puntuale

10.1 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 i parametri sono noti la legge di probabilità è completamente specificata.

Per ottenere informazioni sui parametri non noti 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 da essere *rappresentativo della popolazione*. Molti metodi dell'inferenza statistica sono basati sull'ipotesi di *campioni casuali*.

Definizione 10.1 *Si consideri una popolazione descritta da una variabile aleatoria osservabile X caratterizzata da funzione di distribuzione $F_X(x)$. Il vettore aleatorio $X_1, X_2, \ldots, X_n$ è detto campione casuale di ampiezza n se le variabili aleatorie del vettore sono osservabili, indipendenti e identicamente distribuite (iid) con la stessa legge di probabilità della popolazione (ossia costituiscono delle osservazioni di X). La funzione di distribuzione del campione casuale è:*

$$\begin{aligned} F_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) &= P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) \\ &= P(X_1 \leq x_1)\, P(X_2 \leq x_2) \cdots P(X_n \leq x_n) = \prod_{i=1}^{n} F_X(x_i). \quad (10.1) \end{aligned}$$

Dalla Definizione 10.1 si nota che il campione casuale può essere estratto da una popolazione illimitata oppure da una popolazione finita; si suppone che l'estra-

zione 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 $X_1, X_2, \ldots, X_n$ di ampiezza n estratto dalla popolazione e si cerca di ottenere informazioni sui parametri non noti facendo uso di alcune variabili aleatorie, che sono funzioni misurabili del campione casuale, dette *statistiche* e *stimatori*.

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

Definizione 10.2 *Uno stimatore* $\widehat{\Theta} = t(X_1, X_2, \ldots, X_n)$ *è una funzione misurabile e osservabile del campione casuale* $X_1, X_2, \ldots, X_n$ *i cui valori possono essere usati per stimare un parametro non noto* ϑ *della popolazione. I valori* $\widehat{\vartheta}$ *assunti da tale stimatore sono detti stime del parametro non noto* ϑ.

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

Definizione 10.3 *Sia* $X_1, X_2, \ldots, X_n$ *un campione casuale. La statistica*

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{10.2}$$

è detta media campionaria, mentre la statistica

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{10.3}$$

è detta varianza campionaria.

Proposizione 10.1 *Sia* $X_1, X_2, \ldots, X_n$ *un campione casuale estratto da una popolazione descritta da una variabile aleatoria osservabile* X *caratterizzata da valore medio* $E(X) = \mu$ *finito e varianza* $\mathrm{Var}\,(X) = \sigma^2$ *finita. Risulta:*

$$E(\overline{X}) = \mu, \qquad \mathrm{Var}\,(\overline{X}) = \frac{\sigma^2}{n}\,. \tag{10.4}$$

Dimostrazione Per la proprietà di linearità del valore medio e l'identica distribuzione delle variabili aleatorie che costituiscono il campione, dalla (10.2) si ha:

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

Inoltre, poiché le le variabili aleatorie che costituiscono il campione sono indipendenti ed identicamente distribuite, dalla (10.2) si ottiene:

$$\mathrm{Var}\,(\overline{X}) = \mathrm{Var}\,\Big[\frac{1}{n} \sum_{i=1}^{n} X_i\Big] = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}\,(X_i) = \frac{\sigma^2}{n}\cdot$$

□

La Proposizione 10.1 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 scaturisce che per n sufficientemente grande (ossia per campioni di grande ampiezza) la funzione di distribuzione della media campionaria $\overline{X}$ è approssimativamente normale con valore medio μ e varianza σ^2/n.

Proposizione 10.2 *Sia $X_1, X_2, \ldots, X_n$ un campione casuale estratto da una popolazione descritta da una variabile aleatoria osservabile X caratterizzata da valore medio $E(X) = \mu$, varianza $\mathrm{Var}\,(X) = \sigma^2$ e avente i primi quattro momenti finiti. Risulta:*

$$E(S^2) = \sigma^2, \qquad \mathrm{Var}\,(S^2) = \frac{1}{n}\Big(\mu_4 - \frac{n-3}{n-1}\,\sigma^4\Big), \tag{10.5}$$

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

Dimostrazione Dimostriamo per semplicità soltanto la prima delle (10.5). Osserviamo in primo luogo che

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \sum_{i=1}^{n}\big[X_i - \mu - (\overline{X} - \mu)\big]^2 \\
&= \sum_{i=1}^{n}\Big\{(X_i - \mu)^2 + (\overline{X} - \mu)^2 - 2\,(X_i - \mu)\,(\overline{X} - \mu)\Big\} \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 + n\,(\overline{X} - \mu)^2 - 2\,(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - n\,(\overline{X} - \mu)^2
\end{aligned}$$

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

$$E(S^2) = E\Big[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\Big] = \frac{1}{n-1}\,E\Big[\sum_{i=1}^{n}(X_i - \mu)^2 - n\,(\overline{X} - \mu)^2\Big]$$

$$= \frac{1}{n-1}\,\Big[\sum_{i=1}^{n}\mathrm{Var}(X_i) - n\,\mathrm{Var}\,(\overline{X})\Big] = \frac{1}{n-1}\,\Big[n\,\sigma^2 - n\,\frac{\sigma^2}{n}\Big] = \sigma^2.$$

□

10.2 Metodi per la ricerca di stimatori

Sia $X_1, X_2, \ldots, X_n$ un campione casuale estratto da una popolazione con funzione di probabilità (nel caso discreto) oppure densità di probabilità (nel caso assolutamente continuo) $f(x; \vartheta_1, \vartheta_2, \ldots, \vartheta_k)$ dove $\vartheta_1, \vartheta_2, \ldots, \vartheta_k$ denotano i parametri non noti della popolazione. Lo scopo del decisore, dopo aver osservato

i valori del 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*.

10.2.1 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 10.4 *Si definisce momento campionario r–esimo relativo ai valori osservati $(x_1, x_2, \ldots, x_n)$ del campione casuale il valore*

$$M_r(x_1, x_2, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i^r \qquad (r = 1, 2, \ldots) \tag{10.6}$$

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 $M_1(x_1, x_2, \ldots, x_n)$ coincide con il valore osservato della media campionaria $\overline{X}$, ossia $M_1 = (x_1 + x_2 + \ldots + x_n)/n$.

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

$$E\big(X^r\big) = M_r(x_1, x_2, \ldots, x_n) \qquad (r = 1, 2, \ldots, k). \tag{10.7}$$

I termini alla sinistra di questo sistema di equazioni dipendono dalla legge di probabilità considerata e contengono i parametri non noti della popolazione. Invece, i termini alla destra possono essere calcolati a partire dai dati osservati del campione estratto dalla popolazione. Le incognite del sistema sono i parametri $\vartheta_1, \vartheta_2, \ldots, \vartheta_k$ e sono presenti alla sinistra di questo sistema di equazioni.

Affinché il metodo dei momenti sia utilizzabile occorre che il sistema (10.7) ammetta un'unica soluzione. Le stime dei parametri ottenute con tale metodo, indicate con $\widehat{\vartheta}_1$, $\widehat{\vartheta}_2$, $\ldots$, $\widehat{\vartheta_k}$, dipendono dal campione osservato $(x_1, x_2, \ldots, x_n)$ e quindi al variare dei possibili campioni osservati si ottengono gli stimatori $\widehat{\Theta}_1, \widehat{\Theta}_2, \ldots, \widehat{\Theta}_k$ dei parametri non noti $\vartheta_1, \vartheta_2, \ldots, \vartheta_k$ 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 proponiamo di determinare con il metodo dei momenti lo stimatore del parametro p di una popolazione di Bernoulli descritta da una variabile aleatoria $X \sim \mathcal{B}(1, p)$ con funzione di probabilità

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

Occorre quindi stimare il parametro p. Poiché $E(X) = p$, dalla (10.7) si ha:

$$\widehat{p} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \overline{x}.$$

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

Esempio 10.1 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 $\widehat{p} = 0.5667$. ◊

► (**Popolazione binomiale**) Ci proponiamo di determinare con il metodo dei momenti lo stimatore del parametro p di una popolazione binomiale descritta da una variabile aleatoria $X \sim \mathcal{B}(k, p)$ con funzione di probabilità

$$p_X(x) = \binom{k}{x} p^x (1-p)^{k-x}, \qquad x = 0, 1, \ldots, k \qquad (0 < p < 1).$$

Occorre quindi stimare il parametro p. Poiché $X = Y_1 + Y_2 + \ldots + Y_k$, dove $Y_1, Y_2, \ldots, Y_k$ sono variabili aleatorie indipendenti di Bernoulli, risulta $E(X) = k\,p$ e dalla (10.7) si ha:

$$k\,\widehat{p} = \frac{x_1 + x_2 + \ldots + x_n}{n}, \quad \text{ossia} \quad \widehat{p} = \frac{1}{k}\,\frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\overline{x}}{k}.$$

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

Esempio 10.2 Consideriamo un campione campbinom di ampiezza $n = 30$ contenente come *risultati il numero di successi ottenuti in* $k = 10$ *lanci indipendenti di una moneta*

```
> campbinom<-c(3, 2, 6, 2, 4, 4, 7, 4, 6, 6, 5, 4, 5, 4, 8,
+ 1, 3, 7, 4, 0, 3, 7, 4, 4, 3, 2, 5, 5, 3, 2)
>
> lanci<-10
> stimap<-mean(campbinom)/lanci
> stimap
[1] 0.41
```

la stima del parametro p con il metodo dei momenti è $\widehat{p} = 0.41$. ◊

► (**Popolazione geometrica**) Ci si propone di determinare con il metodo dei momenti lo stimatore del parametro p di una popolazione geometrica descritta da una variabile aleatoria $Y \sim \mathcal{BN}(1, p)$ con funzione di probabilità

$$p_Y(y) = (1-p)^y\,p, \qquad y = 0, 1, \ldots \qquad (0 < p < 1).$$

Poiché $E(Y) = (1-p)/p$, ponendo $\vartheta = (1-p)/p$ dalla (10.7) si ha:

$$\widehat{\vartheta} = \frac{x_1 + x_2 + \ldots + x_n}{n}, \quad \text{ossia} \quad \frac{1-\widehat{p}}{\widehat{p}} = \overline{x}, \quad \text{ossia} \quad \widehat{p} = \frac{1}{1+\overline{x}}.$$

Il metodo dei momenti fornisce quindi come stimatore di $\vartheta = E(Y) = (1-p)/p$ la media campionaria $\overline{X}$.

Esempio 10.3 Consideriamo un campione campgeom di ampiezza 30 contenente come *risultati il numero di fallimenti prima di ottenere il primo successo in lanci ripetuti di una moneta* si ha

```
> campgeom<-c(7, 2, 4, 0, 1, 1, 0, 2, 8, 2, 1, 0, 8, 1, 0,
+ 10, 9, 1, 5, 8, 0, 4, 3, 7,15, 9, 1, 0, 0, 0)
>
> stimap<-1/(1+mean(campgeom))
> stimap
[1] 0.2158273
```

la stima del parametro p con il metodo dei momenti è $\widehat{p} = 0.2158$. ◊

► (**Popolazione geometrica modificata**) Ci proponiamo di determinare con il metodo dei momenti lo stimatore del parametro p di una popolazione geometrica modificata descritta da una variabile aleatoria $X \sim \mathcal{BN}^*(1,p)$ con funzione di probabilità

$$p_X(x) = p\,(1-p)^{x-1}, \qquad x = 1, 2, \ldots \qquad (0 < p < 1).$$

Poiché $E(X) = 1/p$, ponendo $\vartheta = 1/p$ dalla (10.7) si ha:

$$\widehat{\vartheta} = \frac{x_1 + x_2 + \ldots + x_n}{n}, \quad \text{ossia} \quad \frac{1}{\widehat{p}} = \overline{x}, \quad \text{ossia} \quad \widehat{p} = \frac{1}{\overline{x}}.$$

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

Esempio 10.4 Consideriamo un campione campgeommod di ampiezza 30 contenente come *risultati i tempi di attesa per ottenere il primo successo in lanci ripetuti di una moneta*

```
> campgeommod<-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(campgeommod)
> stimap
[1] 0.2752294
```

la stima del parametro p con il metodo dei momenti è $\widehat{p} = 0.2752$. ◊

► (**Popolazione binomiale negativa**) Ci proponiamo di determinare con il metodo dei momenti lo stimatore del parametro p di una popolazione binomiale negativa descritta da una variabile aleatoria $Y \sim \mathcal{BN}(k,p)$, dove $Y = Y_1 + Y_2 +$

$\ldots + Y_k$, essendo $Y_1, Y_2, \ldots, Y_k$ variabili aleatorie indipendenti e identicamente distribuite in modo geometrico. Poiché $E(Y) = k(1-p)/p$, ponendo $\vartheta = k\,(1-p)/p$ dalla (10.7) si ha:

$$\widehat{\vartheta} = \frac{x_1 + x_2 + \ldots + x_n}{n}, \quad \text{ossia} \quad k\,\frac{1-\widehat{p}}{\widehat{p}} = \overline{x}, \quad \text{ossia} \quad \widehat{p} = \frac{k}{k+\overline{x}}.$$

Il metodo dei momenti fornisce quindi come stimatore di $\vartheta = E(Y) = k(1-p)/p$ la media campionaria $\overline{X}$.

Esempio 10.5 Consideriamo un campione campgeom di ampiezza $n = 30$ contenente come *risultati il numero di fallimenti prima di ottenere il quinto successo in lanci ripetuti di una moneta* si ha

```
> campbinneg<-c(2,3,6,20,9,8,13 ,12 ,10 ,21,  8, 13, 18, 21,  5,
+ 8,  5, 26,  4,  8,  6,  3,  3,  1, 10, 13,  3,  7, 10, 17)
>
> k<-5
> stimap<-k/(k+mean(campbinneg))
> stimap
[1] 0.3386005
```

la stima del parametro p con il metodo dei momenti è $\widehat{p} = 0.3386$. ◊

► (**Popolazione binomiale negativa modificata**) Ci proponiamo di determinare con il metodo dei momenti lo stimatore del parametro p di una popolazione binomiale negativa modificata descritta da una variabile aleatoria $X \sim \mathcal{BN}^*(k,p)$, dove $X = Z_1 + Z_2 + \ldots + Z_k$, essendo $Z_1, Z_2, \ldots, Z_k$ variabili aleatorie indipendenti e identicamente distribuite di tipo geometrico modificato. Poiché $E(X) = k/p$, ponendo $\vartheta = k/p$ dalla (10.7) si ha:

$$\widehat{\vartheta} = \frac{x_1 + x_2 + \ldots + x_n}{n}, \quad \text{ossia} \quad \frac{k}{\widehat{p}} = \overline{x}, \quad \text{ossia} \quad \widehat{p} = \frac{k}{\overline{x}}.$$

Il metodo dei momenti fornisce quindi come stimatore del parametro $\vartheta = E(X) = k/p$ la media campionaria $\overline{X}$.

Esempio 10.6 Consideriamo un campione campgeommod di ampiezza 30 contenente come *risultati i tempi di attesa per ottenere il quinto successo in lanci ripetuti di una moneta*

```
> campbinnegmod<-c(20,22,16,13,12,13,28,23,19, 9,19,15,12,14,16,
+ 15, 9,17, 7,15,17, 8,16,20,13,12,14,16,31,14)
>
> k<-5
> stimap<-k/mean(campbinnegmod)
> stimap
[1] 0.3157895
```

la stima del parametro p con il metodo dei momenti è $\widehat{p} = 0.3158$. ◊

► (**Popolazione di Poisson**) Si desidera determinare con il metodo dei momenti lo stimatore del valore medio λ di una popolazione di Poisson descritta da una variabile aleatoria $X \sim \mathcal{P}(\lambda)$ con funzione di probabilità:

$$p_X(x) = \frac{\lambda^x}{x!}\, e^{-\lambda}, \qquad x = 0, 1, \ldots \qquad (\lambda > 0).$$

Occorre quindi stimare il parametro λ. Poiché $E(X) = \lambda$, dalla (10.7) si ha:

$$\widehat{\lambda} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \overline{x}.$$

Il metodo dei momenti fornisce quindi come stimatore del parametro $E(X) = \lambda$ la media campionaria $\overline{X}$.

Esempio 10.7 Consideriamo un campione camppois di ampiezza 30 contenente come *risultati 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 $\widehat{\lambda} = 2.5$. ◊

► (**Popolazione uniforme**) Si desidera determinare con il metodo dei momenti lo stimatore del parametro ϑ di una popolazione uniforme descritta da una variabile aleatoria $X \sim \mathcal{U}(0, \vartheta)$ con funzione densità di probabilità

$$f_X(x) = \frac{1}{\vartheta}, \qquad 0 < x < \vartheta.$$

Occorre quindi stimare il parametro ϑ. Poiché $E(X) = \vartheta/2$, dalla (10.7) si ha:

$$\frac{\widehat{\vartheta}}{2} = \frac{x_1 + x_2 + \ldots + x_n}{n}, \quad \text{ossia} \quad \widehat{\vartheta} = 2\,\overline{x}.$$

Il metodo dei momenti fornisce quindi come stimatore del valore medio $E(X) = \vartheta/2$ la media campionaria $\overline{X}$.

Esempio 10.8 Consideriamo un campione campunif di ampiezza 30 contenente come risultati *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 $\widehat{\vartheta} = 1.4451$. ◊

► (**Popolazione esponenziale**) Si desidera determinare con il metodo dei momenti lo stimatore del valore medio $1/\lambda$ di una popolazione esponenziale descritta da una variabile aleatoria $X \sim \mathcal{E}(\lambda)$ con densità di probabilità

$$f_X(x) = \lambda\, e^{-\lambda\, x}, \qquad x > 0 \qquad (\lambda > 0).$$

Poiché $E(X) = 1/\lambda$, ponendo $\vartheta = 1/\lambda$, dalla (10.7) segue

$$\widehat{\vartheta} = \frac{x_1 + x_2 + \ldots + x_n}{n}, \quad \text{ossia} \quad \widehat{\lambda} = \frac{1}{\overline{x}}.$$

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

Esempio 10.9 Consideriamo un campione campexp di ampiezza 30 contenente come *risultati i tempi di interarrivo (tra arrivi successivi), supposti esponenziali, 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 $\widehat{\vartheta} = 1.8108$. ◊

► (**Popolazione normale**) Si è interessati a determinare con il metodo dei momenti gli stimatori dei parametri μ e σ^2 di una popolazione normale descritta da una variabile aleatoria $X \sim \mathcal{N}(\mu, \sigma)$ di densità di probabilità

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

Occorre quindi stimare due parametri μ e σ^2. Poiché $E(X) = \mu$ e $E(X^2) = \sigma^2 + \mu^2$, dalla (10.7) si ha:

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

da cui si ricava:

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

$$= \frac{1}{n}\sum_{i=1}^{n}\big(x_i - \overline{x}\big)^2 = \frac{(n-1)\,s^2}{n}.$$

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 $(n-1)S^2/n$.

Esempio 10.10 Consideriamo un campione campnorm di ampiezza 30 contenente come *risultati 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 $\widehat{\mu} = 3.0525$ e la stima del parametro σ^2 con il metodo dei momenti è $\widehat{\sigma^2} = 0.02996$. $\Diamond$

10.2.2 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 10.5 *Sia $X_1, X_2, \ldots, X_n$ un campione casuale di ampiezza n estratto dalla popolazione. La funzione di verosimiglianza $L(\vartheta_1, \vartheta_2, \ldots, \vartheta_k) = L(\vartheta_1, \vartheta_2, \ldots, \vartheta_k;\ x_1, x_2, \ldots, x_n)$ del campione osservato $(x_1, x_2, \ldots, x_n)$ 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 $X_1, X_2, \ldots, X_n$, ossia*

$$\begin{aligned} L(\vartheta_1, \vartheta_2, \ldots, \vartheta_k) &= L(\vartheta_1, \vartheta_2, \ldots, \vartheta_k; x_1, x_2, \ldots, x_n) \\ &= f(x_1; \vartheta_1, \vartheta_2, \ldots, \vartheta_k)\ f(x_2; \vartheta_1, \vartheta_2, \ldots, \vartheta_k) \cdots f(x_n; \vartheta_1, \vartheta_2, \ldots, \vartheta_k). \end{aligned} \quad (10.8)$$

Il metodo della massima verosimiglianza consiste nel massimizzare la funzione di verosimiglianza rispetto ai parametri $\vartheta_1, \vartheta_2, \ldots, \vartheta_k$. 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 $(x_1, x_2, \ldots, x_n)$. Pertanto si cercano di determinare i valori $\vartheta_1, \vartheta_2, \ldots, \vartheta_k$ che rendono massima la funzione di verosimiglianza e che quindi offrano, in un certo senso, la migliore spiegazione del campione osservato $(x_1, x_2, \ldots, x_n)$.

I valori di $\vartheta_1, \vartheta_2, \ldots, \vartheta_k$ che massimizzano la funzione di verosimiglianza sono indicati con $\widehat{\vartheta}_1, \widehat{\vartheta}_2, \ldots, \widehat{\vartheta}_k$; essi costituiscono le *stime di massima verosimiglianza* dei parametri non noti $\vartheta_1, \vartheta_2, \ldots, \vartheta_k$ della popolazione. Tali stime

dipendono dal campione osservato $(x_1, x_2, \ldots, x_n)$ e quindi al variare dei possibili campioni osservati si ottengono gli stimatori di massima verosimiglianza $\widehat{\Theta}_1, \widehat{\Theta}_2, \ldots, \widehat{\Theta}_k$ dei parametri non noti $\vartheta_1, \vartheta_2, \ldots, \vartheta_k$ 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 descritta da una variabile aleatoria $X \sim \mathcal{B}(1, p)$ con funzione di probabilità

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

Si ha

$$L(p) = p^{x_1}\,(1-p)^{1-x_1}\; p^{x_2}\,(1-p)^{1-x_2} \cdots p^{x_n}\,(1-p)^{1-x_n}$$

$$= p^{x_1+x_2+\ldots+x_n}\,(1-p)^{n-(x_1+x_2+\ldots+x_n)} \qquad (0 < p < 1),$$

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

$$\log L(p) = \log p \sum_{i=1}^{n} x_i + \Big[n - \sum_{i=1}^{n} x_i\Big] \log(1-p) \qquad (0 < p < 1)$$

da cui si ottiene:

$$\frac{d \log L(p)}{dp} = \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p}\Big[n - \sum_{i=1}^{n} x_i\Big] = \frac{1}{p(1-p)} \sum_{i=1}^{n} x_i - \frac{n}{1-p}$$

$$= \frac{n}{p(1-p)}\Big[\frac{1}{n} \sum_{i=1}^{n} x_i - p\Big].$$

La stima di massima verosimiglianza del parametro p è

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}.$$

Quindi, per una popolazione di Bernoulli lo stimatore di massima verosimiglianza e dei momenti del parametro p è la media campionaria $\overline{X}$. ◊

► (**Popolazione binomiale**) Ci si propone di determinare lo stimatore di massima verosimiglianza del parametro p di una popolazione binomiale descritta da una variabile aleatoria $X \sim \mathcal{B}(k, p)$ con funzione di probabilità

$$p_X(x) = \binom{k}{x} p^x\,(1-p)^{k-x} \qquad (x = 0, 1, \ldots, k).$$

Si ha

$$L(p) = \binom{k}{x_1} p^{x_1}\,(1-p)^{k-x_1} \binom{k}{x_2} p^{x_2}\,(1-p)^{k-x_2} \cdots \binom{k}{x_n} p^{x_n}\,(1-p)^{k-x_n}$$

$$= \binom{k}{x_1}\binom{k}{x_2} \cdots \binom{k}{x_n}\, p^{x_1+x_2+\ldots+x_n}\,(1-p)^{nk-(x_1+x_2+\ldots+x_n)} \qquad (0 < p < 1),$$

dove le x_i possono assumere il valore 0 oppure il valore 1. Procedendo come per la popolazione di Bernoulli, la stima di massima verosimiglianza del parametro kp è

$$k\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{ossia} \qquad \widehat{p} = \frac{1}{k}\,\frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\overline{x}}{k}.$$

Quindi, per una popolazione binomiale lo stimatore di massima verosimiglianza e dei momenti del valore medio $E(X) = k\,p$ è la media campionaria $\overline{X}$. ◊

► (**Popolazione geometrica modificata**) Si è interessati a determinare lo stimatore di massima verosimiglianza del valore medio di una popolazione geometrica modificata descritta da una variabile aleatoria $X \sim \mathcal{BN}^*(1, p)$ con funzione di probabilità

$$p_X(x) = (1-p)^{x-1}\,p, \qquad x = 1, 2, \ldots \qquad (0 < p < 1).$$

Essendo $E(X) = 1/p$, ponendo $\vartheta = 1/p$, si ha

$$\begin{aligned} L(\vartheta) &= \Big(\frac{1}{\vartheta}\Big)^n \Big(1 - \frac{1}{\vartheta}\Big)^{x_1+x_2+\ldots+x_n-n} \\ &= \Big(\frac{1}{\vartheta}\Big)^{x_1+x_2+\ldots+x_n} \big(\vartheta - 1\big)^{x_1+x_2+\ldots+x_n-n} \qquad (\vartheta > 1) \end{aligned}$$

dove le x_i sono numeri interi positivi. Si nota che

$$\log L(\vartheta) = -\log\vartheta \sum_{i=1}^{n} x_i + \log(\vartheta - 1)\,\Big(\sum_{i=1}^{n} x_i - n\Big) \qquad (\vartheta > 1)$$

e quindi si ricava

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

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

$$\widehat{\vartheta} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{ossia} \qquad \widehat{p} = \frac{1}{\overline{x}}.$$

Quindi, per una popolazione geometrica modificata lo stimatore di massima verosimiglianza e dei momenti del valore medio $E(X) = 1/p$ è la media campionaria $\overline{X}$. ◊

► (**Popolazione di Poisson**) Si desidera determinare lo stimatore di massima verosimiglianza del valore medio di una popolazione di Poisson descritta da una variabile aleatoria $X \sim \mathcal{P}(\lambda)$ con funzione di probabilità

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

Essendo $E(X) = \lambda$, si ha

$$L(\lambda) = \frac{\lambda^{x_1+x_2+\ldots+x_n}}{x_1!\, x_2! \cdots x_n!}\, e^{-n\,\lambda} \qquad (\lambda > 0)$$

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

$$\log L(\lambda) = \log \lambda \sum_{i=1}^{n} x_i - n\,\lambda - \log\left[x_1!\, x_2! \cdots x_n!\right] \qquad (\lambda > 0)$$

da cui segue

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

La stima di massima verosimiglianza del parametro λ è

$$\widehat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}.$$

Quindi, per una popolazione di Poisson lo stimatore di massima verosimiglianza e dei momenti di $E(X) = \lambda$ è la media campionaria $\overline{X}$. ◊

► (**Popolazione uniforme**) Si desidera determinare lo stimatore di massima verosimiglianza del parametro ϑ di una popolazione uniforme descritta da una variabile aleatoria $X \sim \mathcal{U}(0, \vartheta)$ con funzione densità di probabilità

$$f_X(x) = \frac{1}{\vartheta}, \qquad 0 < x < \vartheta.$$

Si ha

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

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

$$\widehat{\vartheta} = \max(x_1, x_2, \ldots, x_n).$$

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

Esempio 10.11 Consideriamo un campione campunif di ampiezza 30 contenente come *risultati 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 $\widehat{\vartheta} = 1.856$. ◊

► (**Popolazione esponenziale**) Ci si propone di determinare lo stimatore di massima verosimiglianza del valore medio di una popolazione esponenziale descritta da una variabile aleatoria $X \sim \mathcal{E}(\lambda)$ con funzione densità di probabilità

$$f_X(x) = \lambda\, e^{-\lambda\, x}, \qquad x > 0 \qquad (\lambda > 0).$$

Essendo $E(X) = 1/\lambda$, ponendo $\vartheta = 1/\lambda$, si ha

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

dove le x_i sono positive. Si nota che

$$\log L(\vartheta) = -n\, \log \vartheta - \frac{1}{\vartheta}\sum_{i=1}^{n} x_i \qquad (\vartheta > 0)$$

e quindi si ottiene

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

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

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

Lo stimatore di massima verosimiglianza e dei momenti del valore medio $E(X) = 1/\lambda$ è la media campionaria $\overline{X}$. ◊

► (**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

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

Si ha

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

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

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

e quindi si ha:

$$\frac{\partial \log L(\mu,\,\sigma^2)}{\partial\mu} = \frac{1}{\sigma^2}\,\sum_{i=1}^{n}(x_i-\mu) = \frac{n}{\sigma^2}\,\Big(\frac{1}{n}\,\sum_{i=1}^{n}x_i - \mu\Big)$$

$$\frac{\partial \log L(\mu,\,\sigma^2)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\,\sum_{i=1}^{n}(x_i-\mu)^2 = -\frac{n}{2\sigma^4}\,\Big(\sigma^2 - \frac{1}{n}\,\sum_{i=1}^{n}(x_i-\mu)^2\Big).$$

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

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

Lo stimatore di massima verosimiglianza e dei momenti del valore medio μ è la media campionaria $\overline{X}$. Invece lo stimatore di massima verosimiglianza e dei momenti della varianza σ^2 è $(n-1)S^2/n$. $\Diamond$

10.3 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. Uno stimatore può essere:

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

Definizione 10.6 *Uno stimatore $\widehat{\Theta} = t(X_1, X_2, \ldots, X_n)$ del parametro non noto ϑ della popolazione è detto corretto (non distorto) se e solo se per ogni $\vartheta \in \Theta$ si ha*

$$E(\widehat{\Theta}) = \vartheta, \tag{10.9}$$

ossia se il valore medio dello stimatore $\widehat{\Theta}$ 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 10.1 e 10.2 segue che se $X_1, X_2, \ldots, X_n$ è un campione casuale di ampiezza n estratto da una popolazione caratterizzata da valore medio μ e varianza σ^2, allora $E(\overline{X}) = \mu$ e $E(S^2) = \sigma^2$. Ne segue che *la media campionaria $\overline{X}$ 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 $X_1, X_2, \ldots, X_n$ un campione casuale di ampiezza n estratto da una popolazione di Bernoulli descritta da una variabile aleatoria $X \sim \mathcal{B}(1, p)$ con funzione di probabilità

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

Poiché $E(\overline{X}) = E(X) = p$, la media campionaria $\overline{X}$, 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. ◊

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 $\widehat{\Theta}$ dal parametro non noto ϑ della popolazione.

Definizione 10.7 *Sia $\widehat{\Theta} = t(X_1, X_2, \ldots, X_n)$ uno stimatore del parametro non noto ϑ della popolazione. Si chiama errore quadratico medio la quantità*

$$MSE(\widehat{\Theta}) = E\left[(\widehat{\Theta} - \vartheta)^2\right]. \tag{10.10}$$

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 10.3 *Se $\widehat{\Theta}$ è uno stimatore corretto del parametro ϑ, allora*

$$MSE(\widehat{\Theta}) = E\left\{[\widehat{\Theta} - E(\widehat{\Theta})]^2\right\} = \mathrm{Var}(\widehat{\Theta}). \tag{10.11}$$

La Proposizione 10.3 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 10.8 *Uno stimatore $\widehat{\Theta}$ si dice corretto con varianza uniformemente minima per il parametro non noto ϑ se e solo se per ogni $\vartheta \in \Theta$ risulta*

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

(ii) $\mathrm{Var}(\widehat{\Theta}) \leq \mathrm{Var}(\widehat{\Theta}^\star)$ *per ogni altro stimatore* $\widehat{\Theta}^\star$ *corretto del parametro* ϑ.

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 10.4 *(Disuguaglianza di Cramér–Rao) Sia* $\widehat{\Theta} = t(X_1, X_2, \ldots, X_n)$ *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*

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

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

la varianza dello stimatore $\widehat{\Theta}$ *soddisfa la disuguaglianza*

$$\mathrm{Var}(\widehat{\Theta}) \geq \frac{1}{nE\left\{ \left[\frac{\partial}{\partial\vartheta} \log f(X;\vartheta) \right]^2 \right\}}. \tag{10.12}$$

Si noti che la disuguaglianza di Cramér–Rao (10.12) 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 10.5 *Nelle ipotesi della Proposizione 10.4, se*

$$\mathrm{Var}(\widehat{\Theta}) = \frac{1}{nE\left\{ \left[\frac{\partial}{\partial\vartheta} \log f(X;\vartheta) \right]^2 \right\}}, \tag{10.13}$$

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

► (**Popolazione di Poisson**) Si desidera verificare che $\overline{X}$ è uno stimatore corretto con varianza uniformemente minima del valore medio $E(X) = \lambda$ di una popolazione di Poisson descritta da una variabile aleatoria $X \sim \mathcal{P}(\lambda)$. Tale stimatore è stato determinato sia con il metodo dei momenti che con il metodo della massima verosimiglianza.

La funzione di probabilità è:

$$p_X(x) = \frac{\lambda^x}{x!}\, e^{-\lambda}, \qquad x = 0, 1, \ldots \qquad (\lambda > 0).$$

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

$$\log p(x;\lambda) = -\log x! + x \log\lambda - \lambda$$

e quindi

$$\frac{\partial}{\partial\lambda}\log p(x;\lambda) = \frac{x}{\lambda} - 1 = \frac{x-\lambda}{\lambda}\,.$$

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

$$E\left\{\left[\frac{\partial}{\partial\lambda}\log p(X;\lambda)\right]^2\right\} = E\left[\left(\frac{X-\lambda}{\lambda}\right)^2\right] = \frac{1}{\lambda^2}\,E[(X-\lambda)^2] = \frac{\mathrm{Var}(X)}{\lambda^2} = \frac{1}{\lambda}$$

e quindi

$$\mathrm{Var}(\overline{X}) = \frac{\mathrm{Var}(X)}{n} = \frac{\lambda}{n}\,, \qquad \frac{1}{nE\left\{\left[\frac{\partial}{\partial\lambda}\log p(X;\lambda)\right]^2\right\}} = \frac{\lambda}{n}.$$

Dalla Proposizione 10.5 si ha quindi che $\overline{X}$ è uno stimatore corretto con varianza uniformemente minima del valore medio λ di una popolazione di Poisson. ◊

► (**Popolazione esponenziale**) Si è interessati a verificare che $\overline{X}$ è uno stimatore corretto con varianza uniformemente minima del valore medio $E(X) = 1/\lambda$ di una popolazione esponenziale descritta da una variabile aleatoria $X \sim \mathcal{E}(\lambda)$. 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:

$$f_X(x) = \lambda\, e^{-\lambda\, x}, \qquad x > 0 \qquad (\lambda > 0).$$

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

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

e quindi

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

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

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

e quindi

$$\mathrm{Var}(\overline{X}) = \frac{\mathrm{Var}(X)}{n} = \frac{1}{n\lambda^2}\,, \qquad \frac{1}{nE\left\{\left[\frac{\partial}{\partial\vartheta}\log f(X;\vartheta)\right]^2\right\}} = \frac{\vartheta^2}{n} = \frac{1}{n\lambda^2}.$$

Dalla Proposizione 10.5 segue che $\overline{X}$ è uno stimatore corretto con varianza uniformemente minima del valore medio $1/\lambda$ di una popolazione esponenziale. ◊

► (**Popolazione normale**) Si desidera verificare che $\overline{X}$ è uno stimatore corretto con varianza uniformemente minima del valore medio $E(X) = \mu$ di una popolazione normale descritta sa una variabile aleatoria $X \sim \mathcal{N}(\mu, \sigma)$ 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:

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

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

$$\log f(x;\mu) = -\ \log\bigl(\sigma\sqrt{2\,\pi}\bigr) - \frac{(x-\mu)^2}{2\,\sigma^2}$$

e quindi

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

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

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

e quindi

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}\,, \qquad \frac{1}{nE\left\{\left[\dfrac{\partial}{\partial\mu}\log f(X;\mu)\right]^2\right\}} = \frac{\sigma^2}{n}\cdot$$

Dalla Proposizione 10.5 segue quindi che $\overline{X}$ è uno stimatore corretto con varianza uniformemente minima del valore medio μ di una popolazione normale con varianza nota σ^2. ◊

Occorre sottolineare che *la media campionaria $\overline{X}$ non è sempre uno stimatore corretto con varianza uniformemente minima del valore medio di una specifica 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), …, $(X_1, X_2, \ldots, X_n)$, … di ampiezze crescenti e definiamo una successione $\widehat{\Theta}_1 = t(X_1)$, $\widehat{\Theta}_2 = t(X_1, X_2)$, …, $\widehat{\Theta}_n = t(X_1, X_2, \ldots, X_n)$, … 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 10.9 *Uno stimatore $\widehat{\Theta}_n = t(X_1, X_2, \ldots, X_n)$ 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*

$$\lim_{n\to+\infty} E(\widehat{\Theta}_n) = \vartheta, \tag{10.14}$$

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

► **(Stimatore asintoticamente corretto della varianza di una popolazione)** Si desidera verificare che

$$\widehat{\Theta}_n = \frac{n-1}{n}\, S^2 = \frac{1}{n}\, \sum_{i=1}^{n} \big(X_i - \overline{X}\big)^2$$

è uno stimatore asintoticamente corretto della varianza σ^2 di una popolazione.

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

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

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. ◊

► (**Popolazione di Bernoulli**) Per una popolazione di Bernoulli descritta da una variabile aleatoria $X \sim \mathcal{B}(1,p)$ si desidera verificare che

$$\widehat{\Theta}_n = \frac{n\,\overline{X}+1}{n+2}$$

è uno stimatore asintoticamente corretto del parametro p.

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

$$\lim_{n\to+\infty} E(\widehat{\Theta}_n) = \lim_{n\to+\infty} E\Big(\frac{n\,\overline{X}+1}{n+2}\Big) = \lim_{n\to+\infty} \frac{n\,E(\overline{X})+1}{n+2} = p.$$

◊

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

Definizione 10.10 *Uno stimatore $\widehat{\Theta}_n = t(X_1, X_2, \ldots, X_n)$ del parametro non noto ϑ della popolazione è detto consistente se e solo se per ogni $\varepsilon > 0$ si ha*

$$\lim_{n\to+\infty} P\big(|\widehat{\Theta}_n - \vartheta| < \varepsilon\big) = 1,$$

ossia se e solo se $\widehat{\Theta}_n$ 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 $\overline{X}$ è uno stimatore consistente per μ, ossia per ogni $\varepsilon > 0$ si ha

$$\lim_{n\to+\infty} P\big(|\overline{X}_n - \mu| < \varepsilon\big) = 1,$$

Una condizione sufficiente affinché uno stimatore sia consistente è fornita nella seguente proposizione.

Proposizione 10.6 *Lo stimatore* $\widehat{\Theta}_n = t(X_1, X_2, \ldots, X_n)$ *del parametro non noto* ϑ *della popolazione è consistente se*

i) $\lim_{n\to\infty} E(\widehat{\Theta}_n) = \vartheta$,

ii) $\lim_{n\to+\infty} \mathrm{Var}(\widehat{\Theta}_n) = 0$.

Quindi, una condizione sufficiente affinché lo stimatore sia consistente è che sia asintotiocamente corretto e la sua varianza tende a zero al crescere del campione. Si noti che la Proposizione 10.6 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 descritta da una variabile aleatoria $X \sim \mathcal{U}(0, \vartheta)$ con funzione densità di probabilità

$$f_X(x) = \frac{1}{\vartheta}, \qquad 0 < x < \vartheta.$$

Abbiamo precedentemente mostrato che lo stimatore di ϑ ottenuto con il metodo dei momenti è $\widehat{\Theta} = 2\overline{X}$, mentre quello ottenuto con il metodo della massima verosimiglianza è $\widehat{\Theta} = \max(X_1, X_2, \ldots, X_n)$. 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 $\widehat{\Theta} = 2\overline{X}$ è corretto e consistente. Infatti, si ha

$$E(\widehat{\Theta}) = 2\,E(\overline{X}) = 2\,E(X) = 2\,\frac{\vartheta}{2} = \vartheta,$$
$$\mathrm{Var}(\widehat{\Theta}) = 4\mathrm{Var}(\overline{X}) = 4\,\frac{\mathrm{Var}(X)}{n} = \frac{4}{n}\,\frac{\vartheta^2}{12} = \frac{\vartheta^2}{3n},$$

da cui segue che lo stimatore $\widehat{\Theta} = 2\overline{X}$ del parametro ϑ è corretto e consistente, essendo $\lim_{n\to+\infty} \mathrm{Var}(\widehat{\Theta}) = 0$.

- **Stimatore di** ϑ **con il metodo della massima verosimiglianza.** Vogliamo mostrare che $\widehat{\Theta} = \max(X_1, X_2, \ldots, X_n)$ è asintoticamente corretto e consistente. Osserviamo che per l'indipendenza e l'identica distribuzione delle variabili del campione casuale si ha:

$$P(\widehat{\Theta} \le x) = P(\max(X_1, X_2, \ldots, X_n) \le x) = P(X_1 \le x, X_2 \le x, \ldots, X_n \le x)$$

$$= P(X_1 \le x)\, P(X_2 \le x) \cdots P(X_n \le x) = \begin{cases} 0, & x < 0 \\ \left(x/\vartheta\right)^n, & 0 \le x < \vartheta \\ 1, & x \ge \vartheta. \end{cases}$$

da cui la densità dello stimatore $\widehat{\Theta} = \max(X_1, X_2, \ldots, X_n)$ è

$$f(x) = \begin{cases} n\vartheta^{-n} x^{n-1}, & 0 < x < \vartheta \\ 0, & \text{altrimenti.} \end{cases}$$

Si ricava quindi che

$$E(\widehat{\Theta}) = \int_0^{\vartheta} x\, f(x)\, dx = \frac{n}{n+1}\,\vartheta,$$
$$E(\widehat{\Theta}^2) = \int_0^{\vartheta} x^2\, f(x)\, dx = \frac{n}{n+2}\,\vartheta^2$$
$$\mathrm{Var}(\widehat{\Theta}) = E(\widehat{\Theta}^2) - [E(\widehat{\Theta})]^2 = \frac{n}{(n+2)(n+1)^2}\vartheta^2,$$

da cui segue che lo stimatore $\widehat{\Theta} = \max(X_1, X_2, \ldots, X_n)$ del parametro ϑ è asintoticamente corretto e consistente, essendo $\lim_{n\to+\infty} \mathrm{Var}(\widehat{\Theta}) = 0$.

- **Altro stimatore corretto di ϑ.** Consideriamo infine un altro stimatore corretto di ϑ

$$\widehat{\Theta}_1 = \frac{n+1}{n} \max(X_1, X_2, \ldots, X_n)$$

Tale stimatore gode delle seguenti proprietà:

$$E(\widehat{\Theta}_1) = \frac{n+1}{n}\, E[\max(X_1, X_2, \ldots, X_n)] = \vartheta,$$
$$\mathrm{Var}(\widehat{\Theta}_1) = \frac{(n+1)^2}{n^2}\, \mathrm{Var}[\max(X_1, X_2, \ldots, X_n)] = \frac{1}{n(n+2)}\,\vartheta^2,$$

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 $(n+1)\max(X_1, X_2, \ldots, X_n)/n$ è più efficiente di $\widehat{\Theta} = 2\overline{X}$. ♢

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 10.1 riassumiamo le proprietà di alcuni stimatori precedentemente considerati per il valore medio e per la varianza di alcune popolazioni.

Tabella 10.1: Stimatori del valore medio e della varianza e loro proprietà

X	Metodo dei momenti	Metodo della massima verosimiglianza	Proprietà degli stimatori
Bernoulli $X \sim \mathcal{B}(1,p)$ $E(X) = p$	$\overline{X}$	$\overline{X}$	Stimatore corretto con varianza minima e consistente per p
Binomiale $X \sim \mathcal{B}(k,p)$ $E(X) = k\,p$	$\overline{X}$	$\overline{X}$	Stimatore corretto con varianza minima e consistente per $k\,p$
Geometrica $X \sim \mathcal{BN}(1,p)$ $E(X) = (1-p)/p$	$\overline{X}$	$\overline{X}$	Stimatore corretto con varianza minima e consistente per $(1-p)/p$
Geometrica modificata $X \sim \mathcal{BN}^*(1,p)$ $E(X) = 1/p$	$\overline{X}$	$\overline{X}$	Stimatore corretto con varianza minima e consistente per $1/p$
Binomiale Negativa $X \sim \mathcal{BN}(k,p)$ $E(X) = k(1-p)/p$	$\overline{X}$	$\overline{X}$	Stimatore corretto con varianza minima e consistente per $k(1-p)/p$
Binomiale Negativa modificata $X \sim \mathcal{BN}^*(k,p)$ $E(X) = k/p$	$\overline{X}$	$\overline{X}$	Stimatore corretto con varianza minima e consistente per k/p
Poisson $X \sim \mathcal{P}(\lambda)$ $E(X) = \lambda$	$\overline{X}$	$\overline{X}$	Stimatore corretto con varianza minima e consistente per λ
Uniforme in $(0,\vartheta)$ $X \sim \mathcal{U}(0,\vartheta)$ $E(X) = \vartheta/2$	(1) $\overline{X}$	(2) $\frac{\max(X_1,X_2,\ldots,X_n)}{2}$	(1) Stimatore corretto e consistente per $\vartheta/2$ (2) Stimatore asintoticamente corretto e consistente per $\vartheta/2$
Esponenziale $X \sim \mathcal{E}(\lambda)$ $E(X) = 1/\lambda$	$\overline{X}$	$\overline{X}$	Stimatore corretto con varianza minima e consistente per $1/\lambda$
Normale $X \sim \mathcal{N}(\mu,\sigma)$ $E(X) = \mu$ $\mathrm{Var}(X) = \sigma^2$	(1) $\overline{X}$ (2) $\frac{(n-1)S^2}{n}$	(1) $\overline{X}$ (2) $\frac{(n-1)S^2}{n}$	(1) Stimatore corretto con varianza minima e consistente per μ (2) Stimatore asintoticamente corretto e consistente per σ^2

Capitolo 11

Intervalli di confidenza

11.1 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 $X_1, X_2, \ldots, X_n$ 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 $\underline{C}_n = g_1(X_1, X_2, \ldots, X_n)$ e con $\overline{C}_n = g_2(X_1, X_2, \ldots, X_n)$ 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 $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ risulti $g_1(\mathbf{x}) < g_2(\mathbf{x})$.

Definizione 11.1 *Fissato un coefficiente di confidenza* $1-\alpha \quad (0<\alpha<1)$*, se è possibile scegliere le statistiche* $\underline{C}_n$ *e* $\overline{C}_n$ *in modo tale che*

$$P\big(\underline{C}_n < \vartheta < \overline{C}_n\big) = 1-\alpha,$$

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 $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, allora l'intervallo $\big(g_1(\mathbf{x}), g_2(\mathbf{x})\big)$ 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 con-

fidenza deve essere effettuata dal decisore 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*

$$L(X_1, X_2, \ldots, X_n; 1-\alpha) = \overline{C}_n - \underline{C}_n \tag{11.1}$$

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 $\gamma(X_1, X_2, \ldots, X_n; \vartheta)$ che

- dipende dal campione casuale $X_1, X_2, \ldots, X_n$;
- dipende dal parametro non noto ϑ;
- la sua funzione di distribuzione non contiene il parametro ϑ da stimare.

La *variabile aleatoria di pivot 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:

$$P\big(\alpha_1 < \gamma(X_1, X_2, \ldots, X_n; \vartheta) < \alpha_2\big) = 1-\alpha. \tag{11.2}$$

Se per ogni possibile campione osservato $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ e per ogni $\vartheta \in \Theta$, si riesce a dimostrare che

$$\alpha_1 < \gamma(\mathbf{x}; \vartheta) < \alpha_2 \quad \Longleftrightarrow \quad g_1(\mathbf{x}) < \vartheta < g_2(\mathbf{x})$$

con $g_1(\mathbf{x})$ e $g_2(\mathbf{x})$ dipendenti soltanto dal campione osservato, allora la (11.2) è equivalente a richiedere che

$$P\big(g_1(X_1, X_2, \ldots, X_n) < \vartheta < g_2(X_1, X_2, \ldots, X_n)\big) = 1-\alpha.$$

Denotando con $\underline{C}_n = g_1(X_1, X_2, \ldots, X_n)$ e $\overline{C}_n = g_2(X_1, X_2, \ldots, X_n)$, dalla Definizione 11.1 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 (11.2) è senza dubbio soddisfatta se per ogni campione osservato $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ e per ogni $\vartheta \in \Theta$ risulta che $\gamma(\mathbf{x}; \vartheta)$ è una funzione strettamente monotona in ϑ.

11.2 Popolazione normale

Sia $X_1, X_2, \ldots, X_n$ un campione casuale di ampiezza n estratto da una popolazione normale descritta da una variabile aleatoria $X = \mathcal{N}(\mu, \sigma)$ 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**)
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

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}},$$

che è distribuita normalmente con valore medio nullo e varianza unitaria, ossia è una normale standard. Tale variabile aleatoria dipende dal campione casuale e dal parametro non noto μ (la varianza σ^2 è nota) e la sua legge di probabilità non dipende dal parametro non noto. Quindi, Z_n può essere interpretata come una variabile aleatoria di pivot. Scegliendo nel metodo pivotale $\alpha_1 = -z_{\alpha/2}$ e $\alpha_2 = z_{\alpha/2}$, dove $z_{\alpha/2}$ è tale che

$$P\big(Z_n < -z_{\alpha/2}\big) = P\big(Z_n > z_{\alpha/2}\big) = \frac{\alpha}{2},$$

si ha:

$$P\big(-z_{\alpha/2} < Z_n < z_{\alpha/2}\big) = 1-\alpha. \tag{11.3}$$

Ciò è evidenziato in Figura 11.1 ottenuta tramite il seguente codice:

```
>curve(dnorm(x,mean=0,sd=1),from=-3, to=3,axes=FALSE,ylim=c(0,0.5),
+xlab="",ylab="",main="Densita' 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()
```

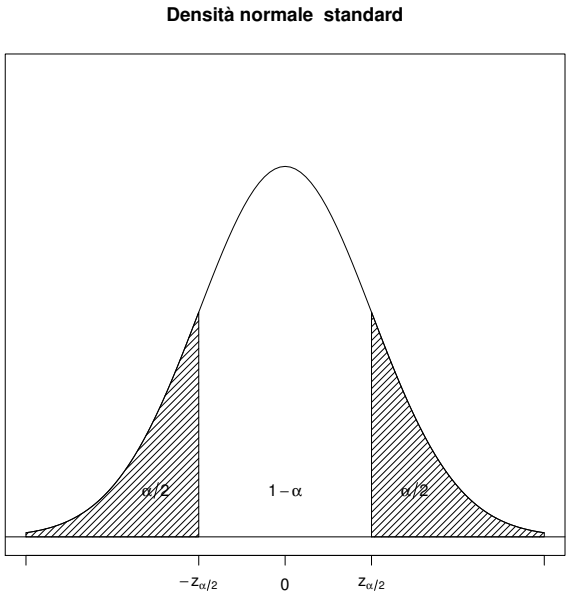

Figura 11.1: Densità normale standard e grado di fiducia $1-\alpha$

Per riempire un grafico come quello di Figura 11.1 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 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 riempendolo di linee inclinate di 45 gradi e equispaziate con densità di 20 per pollice.

Dalla (11.3) si ottiene:

$$P\Big(\overline{X}_n - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\Big) = 1-\alpha.$$

Se poniamo

$$\underline{C}_n = \overline{X}_n - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}, \qquad \overline{C}_n = \overline{X}_n + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\,,$$

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

$$L(X_1, X_2, \ldots, X_n; 1-\alpha) = \overline{C}_n - \underline{C}_n = 2\, z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{11.4}$$

è costante per ogni campione osservato $(x_1, x_2, \ldots, x_n)$. 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 11.1 *Sia $(x_1, x_2, \ldots, x_n)$ 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*

$$\overline{x}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \overline{x}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\,, \tag{11.5}$$

dove

$$\overline{x}_n = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

denota la media campionaria delle n osservazioni.

Esempio 11.1 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 $\overline{x}_{50} = 121.872\, m^2$. 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 11.1, facendo uso della (11.5) 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 $\overline{x}_{50}$ è compresa nell'intervallo. ◊

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 (11.4) si nota che occorre richiedere

$$L(X_1, X_2, \ldots, X_n; 1-\alpha) = \overline{C}_n - \underline{C}_n = 2\, z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq C,$$

da cui segue che

$$n \geq \left(\frac{2\,\sigma}{C}\, z_{\alpha/2} \right)^2.$$

Esempio 11.2 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 $\sigma = 8m^2$ 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 $3\,m^2$. ◊

► (**Intervallo di confidenza per μ con varianza non nota**)

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

$$T_n = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}},$$

dove

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

denota la varianza campionaria. La variabile T_n si può anche così scrivere

$$T_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sqrt{\frac{\sigma^2}{S_n^2}} = \frac{Z_n}{\sqrt{Q_n/(n-1)}}$$

dove

$$Q_n = \frac{(n-1)S_n^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 - \left(\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}\right)^2 \quad (11.6)$$

è distribuita con legge chi-quadrato con $n-1$ gradi di libertà. Dal Teorema 9.4 segue che T_n è distribuita con legge di Student con $n-1$ gradi di libertà. La variabile aleatoria T_n dipende dal campione casuale e dal parametro non noto μ e la sua legge di probabilità non dipende dal parametro non noto. Quindi, T_n si può essere interpretata come una variabile aleatoria di pivot.

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

$$P\big(T_n < -t_{\alpha/2,n-1}\big) = P\big(T_n > t_{\alpha/2,n-1}\big) = \frac{\alpha}{2},$$

si ha:

$$P\big(-t_{\alpha/2,n-1} < T_n < t_{\alpha/2,n-1}\big) = 1-\alpha. \quad (11.7)$$

Ciò è evidenziato in Figura 11.2 ottenuta con $n = 6$ tramite il seguente codice :

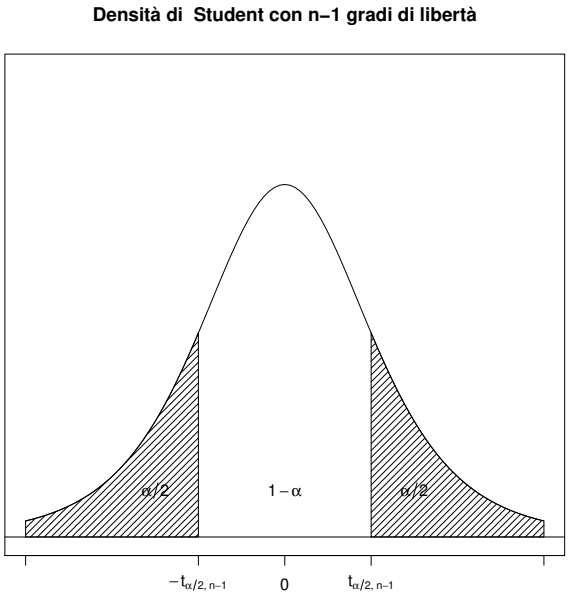

Figura 11.2: Densità di Student con $n-1$ gradi di libertà e grado di fiducia $1-\alpha$

```
>curve(dt(x,df=5),from=-3, to=3,axes=FALSE,ylim=c(0,0.5),xlab="",
+ylab="",main="Densita' di Student con n-1 gradi di liberta' ")
>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)
>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 past(x, y).

Dalla (11.7) si ottiene:

$$P\Big(\overline{X}_n - t_{\alpha/2,n-1}\,\frac{S_n}{\sqrt{n}} < \mu < \overline{X}_n + t_{\alpha/2,n-1}\,\frac{S_n}{\sqrt{n}}\Big) = 1-\alpha.$$

Se poniamo

$$\underline{C}_n = \overline{X}_n - t_{\alpha/2,n-1}\,\frac{S_n}{\sqrt{n}}, \qquad \overline{C}_n = \overline{X}_n + t_{\alpha/2,n-1}\,\frac{S_n}{\sqrt{n}}\,,$$

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 è

$$L(X_1, X_2, \ldots, X_n; 1-\alpha) = \overline{C}_n - \underline{C}_n = 2\,t_{\alpha/2,n-1}\frac{S_n}{\sqrt{n}} \tag{11.8}$$

e si nota che per ogni fissato campione osservato $(x_1, x_2, \ldots, x_n)$ essa cresce al diminuire di α. Quindi, per ogni fissato campione osservato $(x_1, x_2, \ldots, x_n)$ 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 11.2 *Sia $(x_1, x_2, \ldots, x_n)$ 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*

$$\overline{x}_n - t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}} < \mu < \overline{x}_n + t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}} \tag{11.9}$$

dove

$$\overline{x}_n = \frac{x_1 + x_2 + \ldots + x_n}{n}, \qquad s_n = \left\{ \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2 \right\}^{1/2}$$

denotano rispettivamente la media campionaria e la deviazione standard campionaria delle n osservazioni.

Esempio 11.3 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 $\overline{x}_{30} = 10.90\, mg$ e $s_{30} = 0.56\, mg$. 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 11.2, dalla (11.9) 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). ◊

Esempio 11.4 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 $\overline{x}_{100} = 7.2$ con una varianza campionaria $s^2_{100} = 2.25$. 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 $\overline{x}_{200} = 7.2$ e $s^2_{200} = 2.25$, costruire un intervallo di confidenza del 95% per il punteggio medio assegnato dagli studenti.

Nel caso *i)* si ha $\overline{x}_{100} = 7.2$, $s^2_{100} = 2.25$ 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 $\overline{x}_{100} = 7.2$, $s^2_{100} = 2.25$ 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 $\overline{x}_{200} = 7.2$, $s^2_{200} = 2.25$ 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. ◊

Relativamente all'Esempio 11.4, si può determinare il campione di voti tramite la simulazione e ricavare 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 di 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 1 0 1 0 1 0 0 1 1 0 1 0 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 nuovamente il metodo pivotale e consideriamo la variabile aleatoria di pivot

$$V_n = \sum_{i=1}^{n} \Big(\frac{X_i - \mu}{\sigma} \Big)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} \big(X_i - \mu \big)^2$$

Tale variabile aleatoria è distribuita con legge chi–quadrato con n gradi di libertà, essendo costituita dalla somma dei quadrati di n variabili aleatorie normali standard. Inoltre, V_n dipende dal campione casuale e dal parametro non noto σ^2 (essendo il valore medio μ noto) e la sua legge di probabilità non contiene il parametro non noto. Quindi, V_n può essere interpretata come una variabile di pivot.

Scegliendo nel metodo pivotale $\alpha_1 = \chi^2_{1-\alpha/2,n}$ e $\alpha_2 = \chi^2_{\alpha/2,n}$ in maniera tale che

$$P\big(0 < V_n < \chi^2_{1-\alpha/2,n}\big) = P\big(V_n > \chi^2_{\alpha/2,n}\big) = \frac{\alpha}{2}$$

si ha

$$P\big(\chi^2_{1-\alpha/2,n} < V_n < \chi^2_{\alpha/2,n}\big) = 1 - \alpha. \tag{11.10}$$

Ciò è evidenziato in Figura 11.3 ottenuta con $n = 6$ tramite il seguente codice:

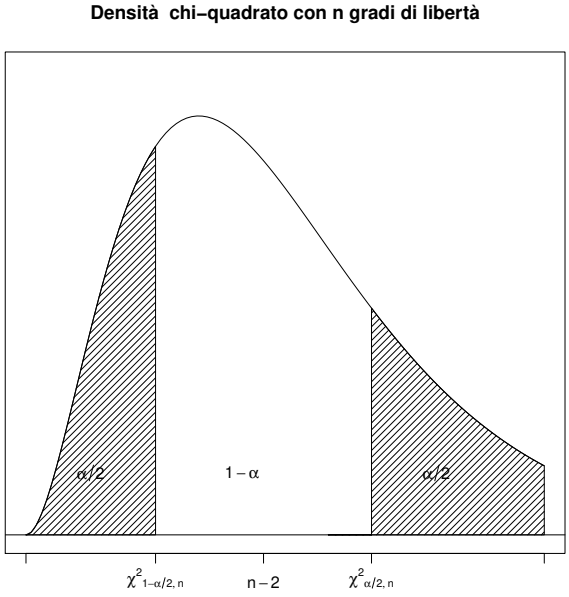

Figura 11.3: Densità chi-quadrato con n gradi di libertà e grado di fiducia $1-\alpha$

```
> curve(dchisq(x,df=6),from=0, to=12,axes=FALSE,ylim=c(0,0.15),
+ xlab="",ylab="",main="Densita' chi-quadrato con n gradi di
    liberta'")
> text(5,0.02,expression(1-alpha))
> axis(1,c(0,3,5.5,8,12),c("",expression({chi^2}[list(1-alpha/2,n)
    ]),
+ expression(n-2),expression({chi^2}[list(alpha/2,n)]),""))
> vals<-seq(0,3,length=100)
> x<-c(0,vals,3,0)
> y<-c(0,dchisq(vals,df=6),0,0)
> polygon(x,y,density=20,angle=45)
> vals<-seq(8,12,length=100)
> x<-c(8,vals,12,7)
> y<-c(0,dchisq(vals,df=6),0,0)
> polygon(x,y,density=20,angle=45)
> abline(h=0)
> text(1.5,0.02,expression(alpha/2))
> text(9.5,0.02,expression(alpha/2))
> box()
```

Poichè V_n si può scrivere in forma alternativa in termini della media campionaria

e della varianza campionaria

$$V_n = \sum_{i=1}^{n}\Big(\frac{X_i - \overline{X} + \overline{X} - \mu}{\sigma}\Big)^2 = \sum_{i=1}^{n}\Big(\frac{X_i - \overline{X}}{\sigma}\Big)^2 + \sum_{i=1}^{n}\Big(\frac{\overline{X} - \mu}{\sigma}\Big)^2$$
$$= \frac{(n-1)\,S_n^2}{\sigma^2} + \Big(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}\Big)^2, \tag{11.11}$$

dalla (11.10) si ottiene

$$P\Big(\chi^2_{1-\alpha/2,n} < \frac{(n-1)\,S_n^2}{\sigma^2} + \Big(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}\Big)^2 < \chi^2_{\alpha/2,n}\Big) = 1-\alpha.$$

o equivalentemente

$$P\left(\frac{(n-1)S_n^2 + n\big(\overline{X}_n - \mu\big)^2}{\chi^2_{\alpha/2,n}} < \sigma^2 < \frac{(n-1)S_n^2 + n\big(\overline{X}_n - \mu\big)^2}{\chi^2_{1-\alpha/2,n}}\right) = 1-\alpha.$$

Se poniamo

$$\underline{C}_n = \frac{(n-1)S_n^2 + n\big(\overline{X}_n - \mu\big)^2}{\chi^2_{\alpha/2,n}}, \qquad \overline{C}_n = \frac{(n-1)S_n^2 + n\big(\overline{X}_n - \mu\big)^2}{\chi^2_{1-\alpha/2,n}}$$

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 11.3 *Sia $(x_1, x_2, \ldots, x_n)$ 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*

$$\frac{(n-1)s_n^2 + n\big(\overline{x}_n - \mu\big)^2}{\chi^2_{\alpha/2,n}} < \sigma^2 < \frac{(n-1)s_n^2 + n\big(\overline{x}_n - \mu\big)^2}{\chi^2_{1-\alpha/2,n}}, \tag{11.12}$$

dove

$$\overline{x}_n = \frac{x_1 + x_2 + \ldots + x_n}{n}, \qquad s_n^2 = \frac{1}{n-1}\sum_{i+1}^{n}(x_i - \overline{x}_n)^2.$$

denotano rispettivamente la media campionaria e la varianza campionaria delle n osservazioni.

Esempio 11.5 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 $\overline{x}_{12} = 70.49\,gr$ e $s^2_{12} = 24.36\ gr^2$. Supponendo che il peso sia distribuito normalmente con valore medio $\mu = 70\,gr$ 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 $\chi^2_{1-\alpha/2,n} = \chi^2_{0.975,12}$ possono essere ottenuti tramite R. Facendo riferimento alla Figura 11.3 e ricordando la (11.12) 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 $\chi^2_{1-\alpha/2,n} = \chi^2_{0.975,12} = 4.404$ e $\chi^2_{\alpha/2,n} = \chi^2_{0.025,12} = 23.337$. La stima dell'intervallo di confidenza di grado $1-\alpha = 0.95$ per la varianza della popolazione normale è quindi (11.61, 61.51). ◊

► (**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

$$Q_n = \frac{(n-1)\,S_n^2}{\sigma^2} = \frac{1}{\sigma^2}\ \sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Dalla (11.11) segue che la variabile aleatoria Q_n è distribuita con legge chi-quadrato con $n-1$ gradi di libertà. La variabile Q_n dipende dal campione casuale e dal parametro non noto σ^2 e la sua legge di probabilità non contiene il parametro non noto. Quindi, Q_n si può interpretare come una variabile di pivot.

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

$$P\big(0 < Q_n < \chi^2_{1-\alpha/2,n-1}\big) = P\big(Q_n > \chi^2_{\alpha/2,n-1}\big) = \frac{\alpha}{2}$$

si ha

$$P\big(\chi^2_{1-\alpha/2,n-1} < Q_n < \chi^2_{\alpha/2,n-1}\big) = 1-\alpha. \tag{11.13}$$

Ciò è evidenziato in Figura 11.4 ottenuta con $n = 7$ tramite il seguente codice:

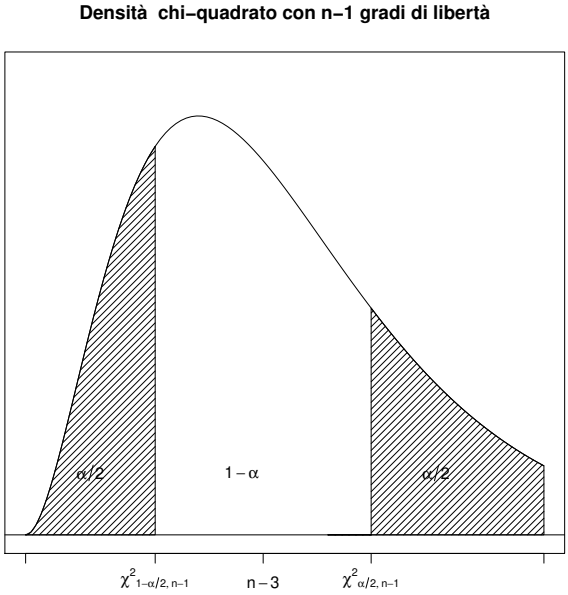

Figura 11.4: Densità chi-quadrato con $n-1$ gradi di libertà e grado di fiducia $1-\alpha$

```
> curve(dchisq(x,df=6),from=0, to=12,axes=FALSE,ylim=c(0,0.15),
+ xlab="",ylab="",main="Densita'  chi-quadrato con n-1 gradi di
    liberta'")
> text(5,0.02,expression(1-alpha))
> axis(1,c(0,3,5.5,8,12),c("",expression({chi^2}[list(1-alpha/2,n
    -1)]),
+ expression(n-3),expression({chi^2}[list(alpha/2,n-1)]),""))
> vals<-seq(0,3,length=100)
> x<-c(0,vals,3,0)
> y<-c(0,dchisq(vals,df=6),0,0)
> polygon(x,y,density=20,angle=45)
> vals<-seq(8,12,length=100)
> x<-c(8,vals,12,7)
> y<-c(0,dchisq(vals,df=6),0,0)
> polygon(x,y,density=20,angle=45)
> abline(h=0)
> text(1.5,0.02,expression(alpha/2))
> text(9.5,0.02,expression(alpha/2))
> box()
```

Dalla (11.13) si ottiene

$$P\Big(\chi^2_{1-\alpha/2,n-1} < \frac{(n-1)\,S_n^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}\Big) = 1-\alpha,$$

che è equivalente a richiedere che

$$P\Big(\frac{(n-1)\,S_n^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)\,S_n^2}{\chi^2_{1-\alpha/2,n-1}}\Big) = 1-\alpha.$$

Se poniamo

$$\underline{C}_n = \frac{(n-1)\,S_n^2}{\chi^2_{\alpha/2,n-1}}\,, \qquad \overline{C}_n = \frac{(n-1)\,S_n^2}{\chi^2_{1-\alpha/2,n-1}}\,,$$

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 11.4 *Sia $(x_1, x_2, \ldots, x_n)$ 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*

$$\frac{(n-1)\,s_n^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)\,s_n^2}{\chi^2_{1-\alpha/2,n-1}}\,, \tag{11.14}$$

dove

$$s_n^2 = \frac{1}{n-1}\sum_{i+1}^{n}(x_i - \overline{x}_n)^2$$

denota la varianza campionaria delle n osservazioni.

Esempio 11.6 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 $\chi^2_{1-\alpha/2,n-1} = \chi^2_{0.975,19}$ e $\chi^2_{\alpha/2,n-1} = \chi^2_{0.025,19}$ possono essere ottenuti tramite R. Infatti, osservando la Figura 11.4 e ricordando la (11.14) 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 $\chi^2_{1-\alpha/2,n-1} = \chi^2_{0.975,19} = 8.907$ e $\chi^2_{\alpha/2,n-1} = \chi^2_{0.025,19} = 32.852$. 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 determinazione di 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.

Nella Tabella 11.1 riassumiamo gli intervalli di confidenza di grado $1-\alpha$ per la stima intervallare del valore medio e della varianza di una popolazione descritta da una variabile aleatoria normale $X \sim \mathcal{N}(\mu, \sigma)$.

Tabella 11.1: Intervalli di confidenza di grado $1-\alpha$ per una popolazione normale

Intervallo di confidenza per μ con varianza σ^2 nota	$\overline{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ $z_{\alpha/2}$ si calcola con $\mathsf{qnorm}(1-\mathsf{alpha}/2, \mathsf{mean}=0, \mathsf{sd}=1)$
Intervallo di confidenza per μ con varianza non nota	$\overline{x}_n - t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}} < \mu < \overline{x}_n + t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}$ $t_{\alpha/2,n-1}$ si calcola con $\mathsf{qt}(1-\mathsf{alpha}/2, \mathsf{df}=\mathsf{n}-1))$
Intervallo di confidenza per σ^2 con valore medio μ noto	$\frac{(n-1)s_n^2 + n\left(\overline{x}_n-\mu\right)^2}{\chi^2_{\alpha/2,n}} < \sigma^2 < \frac{(n-1)s_n^2 + n\left(\overline{x}_n-\mu\right)^2}{\chi^2_{1-\alpha/2,n}}$ $\chi^2_{1-\alpha/2,n}$ si calcola con $\mathsf{qchisq}(\mathsf{alpha}/2, \mathsf{df}=\mathsf{n})$ $\chi^2_{\alpha/2,n}$ si calcola con $\mathsf{qchisq}(1-\mathsf{alpha}/2, \mathsf{df}=\mathsf{n})$
Intervallo di confidenza per σ^2 con valore medio non noto	$\frac{(n-1)\,s_n^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)\,s_n^2}{\chi^2_{1-\alpha/2,n-1}}$ $\chi^2_{1-\alpha/2,n-1}$ si calcola con $\mathsf{qchisq}(\mathsf{alpha}/2, \mathsf{df}=\mathsf{n}-1)$ $\chi^2_{\alpha/2,n-1}$ si calcola con $\mathsf{qchisq}(1-\mathsf{alpha}/2, \mathsf{df}=\mathsf{n}-1)$

Capitolo 12

Intervalli di fiducia approssimati

Ci proponiamo di costruire degli intervalli di confidenza approssimati per campioni di dimensioni elevate utilizzando il teorema centrale di convergenza. Inoltre, desideriamo analizzare alcuni problemi in cui è richiesto il confronto tra i valori medi di due differenti popolazioni; esamineremo il caso di popolazioni normali, di popolazioni di Bernoulli e di popolazioni di Poisson.

12.1 Intervalli di confidenza: grandi campioni

I metodi per la ricerca degli intervalli di confidenza per una popolazione normale, considerati nel Cap. 11, non dipendono dalla dimensione del campione osservato. Se invece la dimensione del campione è elevata ($n \geq 30$) è possibile utilizzare il *teorema centrale di convergenza* per determinare un intervallo di confidenza di grado $1-\alpha$ per il parametro non noto ϑ di una popolazione. Infatti, se X denota la variabile aleatoria che descrive la popolazione con $E(X) = \mu$ e $\mathrm{Var}(X) = \sigma^2$ (supposti entrambi finiti) e con $(X_1, X_2, \ldots, X_n)$ il campione casuale, il teorema centrale di convergenza afferma che la variabile aleatoria

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z,$$

converge in distribuzione ad una variabile aleatoria normale standard.

Se il valore medio $E(X) = \mu$ e $\mathrm{Var}(X) = \sigma^2$ della popolazione dipendono da un parametro non noto ϑ della popolazione, si nota la variabile aleatoria Z_n può essere interpretata come una variabile aleatoria di pivot poiché:

- dipende dal campione casuale $X_1, X_2, \ldots, X_n$;

- dipende dal parametro non noto ϑ della popolazione attraverso il valore medio $E(X) = \mu$ e la varianza $\mathrm{Var}(X) = \sigma^2$;

- per grandi campioni la sua funzione di distribuzione è approssimativamente normale standard e quindi non contiene il parametro ϑ da stimare.

Pertanto, per campioni di ampiezza elevata possiamo applicare il *metodo pivotale in forma approssimata* richiedendo che la (11.3) valga in forma approssimata, ossia

$$P\Big(-z_{\alpha/2} < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\Big) \simeq 1 - \alpha.$$

come evidenziato in Figura 12.1.

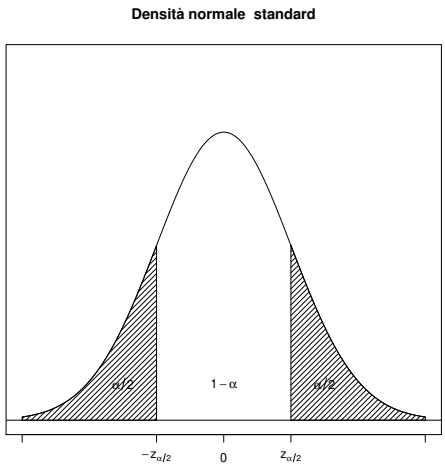

Figura 12.1: Densità normale standard e grado di fiducia $1 - \alpha$

Quando la dimensione del campione è elevata, utilizzeremo il metodo pivotale in forma approssimata nei seguenti casi:

- intervallo di confidenza per il parametro p di una popolazione di Bernoulli;
- intervallo di confidenza per il parametro p di una popolazione binomiale;
- intervallo di confidenza per il parametro p di una popolazione geometrica modificata;
- intervallo di confidenza per il parametro λ di una popolazione di Poisson;
- intervallo di confidenza per il parametro ϑ di una popolazione uniforme;
- intervallo di confidenza per il parametro λ di una popolazione esponenziale.

L'analisi statistica spesso richiede di confrontare parametri di popolazioni differenti, come le medie o le varianze. Utilizzeremo il metodo pivotale per determinare

- un intervallo di confidenza per la differenza tra i valori medi di due popolazioni normali;
- un intervallo di confidenza per la differenza tra i valori medi di due popolazioni di Bernoulli;
- un intervallo di confidenza per la differenza tra i valori medi di due popolazioni di Poisson.

► (**Intervallo di confidenza per il parametro p di una popolazione di Bernoulli**)

Consideriamo una popolazione di Bernoulli descritta da una variabile aleatoria $X \sim \mathcal{B}(1,p)$ con funzione di probabilità

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

Il valore medio di una variabile aleatoria di Bernoulli è $E(X) = p$ e la varianza è $\mathrm{Var}(X) = p\,(1-p)$ ed entrambi dipendono dal parametro non noto p. Ricaviamo che

$$E(\overline{X}_n) = E(X) = p, \qquad \mathrm{Var}(\overline{X}_n) = \frac{\mathrm{Var}(X)}{n} = \frac{p\,(1-p)}{n}.$$

Applicando il teorema centrale di convergenza si ha che la variabile aleatoria

$$\frac{\overline{X}_n - p}{\sqrt{p(1-p)}/\sqrt{n}} = \sqrt{n}\,\frac{\overline{X}_n - p}{\sqrt{p(1-p)}}$$

converge in distribuzione ad una variabile aleatoria normale standard. Per campioni sufficientemente numerosi, l'intervallo di confidenza di grado $1-\alpha$ per il parametro p può essere determinato richiedendo che

$$P\Big(-z_{\alpha/2} < \frac{\sqrt{n}\,(\overline{X}_n - p)}{\sqrt{p(1-p)}} < z_{\alpha/2}\Big) \simeq 1-\alpha. \tag{12.1}$$

La disuguaglianza

$$-z_{\alpha/2} < \frac{\sqrt{n}\,(\overline{x}_n - p)}{\sqrt{p(1-p)}} < z_{\alpha/2}$$

è equivalente a

$$\Big[\frac{\sqrt{n}\,(\overline{x}_n - p)}{\sqrt{p(1-p)}}\Big]^2 < z^2_{\alpha/2},$$

che conduce alla disuguaglianza di secondo grado in p

$$p^2\left(n + z^2_{\alpha/2}\right) - p\left(2\,n\,\overline{x}_n + z^2_{\alpha/2}\right) + n\,\overline{x}^2_n < 0. \tag{12.2}$$

Essendo il coefficiente di p^2 positivo, le soluzioni della disuguaglianza (12.2) sono interne all'intervallo delle radici della corrispondente equazione di secondo grado, ossia $\underline{c}_n < p < \overline{c}_n$.

Il sistema R mette a disposizione la funzione polyroot(c($a_0, a_1, \ldots, a_{n-1}, a_n$)) per calcolare le radici reali e complesse di un'equazione $a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 = 0$. In polyroot(c($a_0, a_1, \ldots, a_{n-1}, a_n$)) i coefficienti del polinomio debbono essere inseriti in ordine crescente rispetto alle potenze del polinomio. Se si denota con

$$a_2 = n + z^2_{\alpha/2}, \qquad a_1 = -\big(2\, n\, \overline{x}_n + z^2_{\alpha/2}\big), \qquad a_0 = n\, \overline{x}^2_n,$$

le radici dell'equazione $a_2 p^2 + a_1 p + a_0 = 0$ possono essere calcolate utilizzando polyroot(c(a_0, a_1, a_2)).

Esempio 12.1 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)
```

Il metodo dei momenti e della massima verosimiglianza hanno fornito come stima del parametro p la media campionaria $\overline{x}_n$. Per il campione considerato abbiamo mostrato che

```
> stimap<-mean(campbern)
> stimap
[1] 0.5666667
```

la stima del parametro p con il metodo dei momenti e con il metodo della massima verosimiglianza è $\widehat{p} = 0.5667$. Vogliamo ora determinare un intervallo di confidenza di grado $1 - \alpha = 0.95$ per il parametro p.

```
> alpha<-1-0.95
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.959964
> zalpha<-qnorm(1-alpha/2,mean=0,sd=1)
> n<-length(campbern)
> a2<-n+zalpha^2
> a1<- -(2*n*mean(campbern)+zalpha**2)
> a0<-n*(mean(campbern))^2
> polyroot(c(a0,a1,a2))
[1] 0.3919731+0i 0.7262251-0i
```

Una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per il parametro p è $(0.3919731, 0.7262251)$. Si nota che la stima puntuale di p, ossia $\widehat{p} = 0.5667$ è contenuta nell'intervallo. ◊

Esempio 12.2 Una ditta farmaceutica è interessata a stabilire l'efficacia di un nuovo farmaco per curare una data malattia. Da un'indagine condotta su 900 pazienti affetti da questa malattia trova che il farmaco è efficace in 740 casi. Sulla base di questi dati si vuole determinare una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la probabilità p che il farmaco sia efficace per l'intera popolazione.

Possiamo supporre che la popolazione sia distribuita secondo Bernoulli, con p che denota la probabilità che il farmaco sia efficace. Il campione è di ampiezza

$n = 900$, dove 900 rappresenta il numero di pazienti esaminati. Poiché per 740 pazienti il farmaco è stato efficace, si ha $\overline{x}_{900} = (x_1 + x_2 + \ldots + x_{900})/900 = 740/900 = 0.822$ (*stima puntuale* di p). Inoltre, essendo $\alpha = 0.05$, si ha $\alpha/2 = 0.025$. Utilizzando R, a partire dalla (12.2) si ha

```
> alpha<-1-0.95
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.959964
> zalpha<-qnorm(1-alpha/2,mean=0,sd=1)
> n<-900
> medcamp<-740/900
> medcamp
[1] 0.8222222
>
> a2<-n+zalpha^2
> a1<- -(2*n*medcamp+zalpha**2)
> a0<-n*medcamp^2
> polyroot(c(a0,a1,a2))
[1] 0.7958901+0i 0.8458153-0i
```

da cui segue che $z_{\alpha/2} = z_{0.025} = 1.96$ e una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per p è $(0.796, 0.846)$. Si nota che la stima puntuale della probabilità con cui il farmaco è efficace per l'intera popolazione, ossia $\widehat{p} = 0.822$ è compresa nell'intervallo. ◊

Esempio 12.3 Un ente di ricerca demoscopica è interessato all'opinione di 100 elettori su una proposta politica. Avendo ottenuto 47 risposte favorevoli desidera determinare l'intervallo di confidenza per la proporzione di risposte favorevoli nella popolazione con un grado di confidenza $1 - \alpha = 0.97$.

Sulla base delle $n = 100$ osservazioni campionarie, la stima per la proporzione di persone che hanno un'opinione favorevole alla proposta politica è $\overline{x}_{100} = 47/100 = 0.47$ (*stima puntuale* di p). Inoltre, $\alpha = 0.03$ e $\alpha/2 = 0.015$. Mediante R, si ha:

```
> alpha<-1-0.97
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 2.17009
> zalpha<-qnorm(1-alpha/2,mean=0,sd=1)
> n<-100
> medcamp<-47/100
>
> a2<-n+zalpha^2
> a1<- -(2*n*medcamp+zalpha**2)
> a0<-n*medcamp^2
> polyroot(c(a0,a1,a2))
[1] 0.3654952-0i 0.5772033+0i
```

da cui segue che $z_{\alpha/2} = z_{0.015} = 2.17$ e una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.97$ per p è $(0.3655, 0.5772)$. Si deduce che la stima della probabilità di opinione favorevole alla proposta politica per l'intera popolazione è piuttosto bassa. ◊

⋆ Metodo alternativo

Possiamo calcolare esplicitamente le radici dell'equazione di secondo grado in p

$$\underline{c}_n, \overline{c}_n = \frac{2n\overline{x}_n + z^2_{\alpha/2} \pm z_{\alpha/2}\sqrt{z^2_{\alpha/2} + 4n\overline{x}_n(1-\overline{x}_n)}}{2(n + z^2_{\alpha/2})}$$

Se nelle stime del limite inferiore e superiore dell'intervallo di confidenza per p trascuriamo i termini che tendono a zero più rapidamente di $1/\sqrt{n}$, si ottiene

$$\frac{2\overline{x}_n + \dfrac{z^2_{\alpha/2}}{n} \pm z_{\alpha/2}\sqrt{\dfrac{z^2_{\alpha/2}}{n^2} + \dfrac{4\overline{x}_n(1-\overline{x}_n)}{n}}}{2\Big(1 + \dfrac{z^2_{\alpha/2}}{n}\Big)} \simeq \overline{x}_n \pm z_{\alpha/2}\sqrt{\frac{\overline{x}_n(1-\overline{x}_n)}{n}}.$$

Sussiste quindi il seguente risultato:

Proposizione 12.1 *Sia $(x_1, x_2, \ldots, x_n)$ un campione osservato di ampiezza n estratto da una popolazione bernoulliana di parametro p. Se la dimensione del campione è elevata, una stima approssimata dell'intervallo di confidenza di grado $1-\alpha$ per p è*

$$\overline{x}_n - z_{\alpha/2}\sqrt{\frac{\overline{x}_n(1-\overline{x}_n)}{n}} < p < \overline{x}_n + z_{\alpha/2}\sqrt{\frac{\overline{x}_n(1-\overline{x}_n)}{n}}. \tag{12.3}$$

dove $\overline{x}_n$ denota la media campionaria.

Relativamente all'Esempio 12.2, dalla Proposizione 12.1 si ottiene l'approssimazione:

```
> medcamp-zalpha*sqrt(medcamp*(1-medcamp)/n)
[1] 0.7972441
> medcamp+zalpha*sqrt(medcamp*(1-medcamp)/n)
[1] 0.8472004
```

mentre per l'Esempio 12.3, dalla Proposizione 12.1 si ottiene:

```
> medcamp-zalpha*sqrt(medcamp*(1-medcamp)/n)
[1] 0.3715823
> medcamp+zalpha*sqrt(medcamp*(1-medcamp)/n)
[1] 0.5884177
```

▶ (**Intervallo di confidenza per il parametro p di una popolazione binomiale**)

Consideriamo una popolazione binomiale descritta da una variabile aleatoria $X \sim \mathcal{B}(k, p)$ con funzione di probabilità

$$p_X(x) = \binom{k}{x} p^x (1-p)^{k-x}, \qquad x = 0, 1, \ldots, k \qquad (0 < p < 1).$$

Il valore medio della variabile aleatoria binomiale è $E(X) = k\,p$ e la varianza $\mathrm{Var}(X) = k\,p\,(1-p)$ ed entrambi dipendono dal parametro non noto p. Ricaviamo che

$$E(\overline{X}_n) = E(X) = k\,p, \qquad \mathrm{Var}(\overline{X}_n) = \frac{\mathrm{Var}(X)}{n} = \frac{k\,p\,(1-p)}{n}.$$

Applicando il teorema centrale di convergenza si ha che la variabile aleatoria

$$\frac{\overline{X}_n - k\,p}{\sqrt{k\,p\,(1-p)/n}} = \sqrt{n}\,\frac{\overline{X}_n - k\,p}{\sqrt{k\,p\,(1-p)}}$$

converge in distribuzione ad una variabile aleatoria normale standard. Per campioni sufficientemente numerosi, l'intervallo di confidenza di grado $1-\alpha$ per il parametro p può essere determinato richiedendo che

$$P\Big(-z_{\alpha/2} < \sqrt{n}\,\frac{\overline{X}_n - k\,p}{\sqrt{k\,p\,(1-p)}} < z_{\alpha/2}\Big) \simeq 1-\alpha. \tag{12.4}$$

La disuguaglianza

$$-z_{\alpha/2} < \sqrt{n}\,\frac{\overline{x}_n - k\,p}{\sqrt{k\,p\,(1-p)}} < z_{\alpha/2},$$

è equivalente a

$$n\,\frac{(\overline{x}_n - k\,p)^2}{k\,p\,(1-p)} < z^2_{\alpha/2},$$

che conduce alla disuguaglianza di secondo grado in p

$$k\,(n\,k + z^2_{\alpha/2})\,p^2 - k\,(2\,n\,\overline{x}_n + z^2_{\alpha/2})\,p + n\,\overline{x}^2_n < 0.$$

Si nota che quando $k=1$ si ottiene la disuguaglianza di secondo grado in p per la popolazione di Bernoulli. Essendo il coefficiente di p^2 positivo, le soluzioni della disuguaglianza sono interne all'intervallo delle radici della corrispondente equazione di secondo grado, ossia $\underline{c}_n < p < \overline{c}_n$.

Se si denota con

$$a_2 = k\,(n\,k + z^2_{\alpha/2}), \qquad a_1 = -k\,(2\,n\,\overline{x}_n + z^2_{\alpha/2}), \qquad a_0 = n\,\overline{x}^2_n,$$

le radici dell'equazione $a_2p^2 + a_1p + a_0 = 0$ si possono calcolare utilizzando la funzione polyroot(c(a_0, a_1, a_2)).

Esempio 12.4 Consideriamo un campione campbinom di ampiezza $n=30$ contenente come *risultati il numero di successi ottenuti in* $k=10$ *lanci indipendenti di una moneta*

```
> campbinom<-c(3, 2, 6, 2, 4, 4, 7, 4, 6, 6, 5, 4, 5, 4, 8,
+ 1, 3, 7, 4, 0, 3, 7, 4, 4, 3, 2, 5, 5, 3, 2)
```

Il metodo dei momenti e della massima verosimiglianza forniva $\overline{x}_n/k$ come stima puntuale del parametro p, dove $k=10$ (numero di lanci della moneta) e $n=30$ è la dimensione del campione. Abbiamo precedentemente mostrato che per il campione in esame risulta

```
> lanci<-10
> stimap<-mean(campbinom)/lanci
> stimap
[1] 0.41
```

che la stima del parametro p con il metodo dei momenti e con il metodo della massima verosimiglianza è $\widehat{p} = 0.41$. Vogliamo ora determinare un intervallo di confidenza con di grado di confidenza $1 - \alpha = 0.95$ per il parametro p.

```
> alpha<-1-0.95
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.959964
> zalpha<-qnorm(1-alpha/2,mean=0,sd=1)
> n<-length(campbinom)
> a2<-lanci*(n*lanci+zalpha^2)
> a1<--lanci*(2*n*mean(campbinom)+zalpha^2)
> a0=n*(mean(campbinom))^2
> polyroot(c(a0,a1,a2))
[1] 0.3558239-0i 0.4664518+0i
```

che mostra che la stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per il parametro p è $(0.3558239, 0.4664518)$. Si nota che la stima puntuale del parametro p, ossia $\widehat{p} = 0.41$, è contenuta nell'intervallo. $\Diamond$

► (**Intervallo di confidenza per il parametro p di una popolazione geometrica modificata**)

Consideriamo una popolazione geometrica modificata descritta da una variabile aleatoria $X \sim \mathcal{BN}^*(1, p)$ caratterizzata da funzione di probabilità

$$p_X(x) = p\,(1-p)^{x-1}, \qquad x = 1, 2, \ldots \qquad (0 < p < 1)$$

Il valore medio di una variabile aleatoria geometrica modificata è $E(X) = 1/p$ e la varianza è $\mathrm{Var}(X) = (1-p)/p^2$ ed entrambi dipendono dal parametro non noto p. Ricaviamo che

$$E(\overline{X}_n) = \frac{1}{p}, \qquad \mathrm{Var}(\overline{X}_n) = \frac{(1-p)}{n\,p^2}.$$

Applicando il teorema centrale di convergenza si ha che la variabile aleatoria

$$\frac{\overline{X}_n - 1/p}{\sqrt{(1-p)/(n\,p^2)}} = \sqrt{n}\,\frac{p\overline{X}_n - 1}{\sqrt{1-p}}$$

converge in distribuzione ad una variabile aleatoria normale standard. Per campioni sufficientemente numerosi, l'intervallo di confidenza di grado $1 - \alpha$ per il parametro p può essere determinato richiedendo che

$$P\Big(-z_{\alpha/2} < \sqrt{n}\,\frac{p\overline{X}_n - 1}{\sqrt{1-p}} < z_{\alpha/2}\Big) \simeq 1 - \alpha. \tag{12.5}$$

La disuguaglianza

$$-z_{\alpha/2} < \sqrt{n}\,\frac{p\overline{x}_n - 1}{\sqrt{1-p}} < z_{\alpha/2}$$

è equivalente a

$$\left[\sqrt{n}\,\frac{p\overline{x}_n - 1}{\sqrt{1-p}}\right]^2 < z^2_{\alpha/2},$$

che conduce alla disuguaglianza di secondo grado in p

$$n\,\overline{x}_n^2\,p^2 - p\,(2\,n\,\overline{x}_n - z_{\alpha/2}^2) + n - z_{\alpha/2}^2 < 0. \tag{12.6}$$

Essendo il coefficiente di p^2 positivo, le soluzioni della disuguaglianza (12.6) sono interne all'intervallo delle radici della corrispondente equazione di secondo grado, ossia $\underline{c}_n < p < \overline{c}_n$. Se si denota con

$$a_2 = n\overline{x}_n^2, \qquad a_1 = -(2n\overline{x}_n - z_{\alpha/2}^2), \qquad a_0 = n - z_{\alpha/2}^2,$$

le radici dell'equazione $a_2p^2 + a_1p + a_0 = 0$ possono essere calcolate mediante $\mathsf{polyroot}(\mathsf{c}(\mathsf{a}_0, \mathsf{a}_1, \mathsf{a}_2))$.

Esempio 12.5 In una produzione di aghi con una macchina automatica, vengono scartati quelli la cui lunghezza è inferiore a 2 cm. Numerando gli aghi prodotti, denotiamo con X la variabile aleatoria che descrive il numero associato al primo ago imperfetto prodotto; la distribuzione di X è geometrica modificata di parametro p, dove p rappresenta la probabilità che l'ago sia imperfetto in una singola produzione. Se si effettuano 100 osservazioni di X, si nota che $\overline{x}_{100} = 10.5$. Determinare un intervallo di confidenza per il parametro p con un grado di confidenza $1 - \alpha = 0.96$.

Nel nostro caso $n = 100$, $\overline{x}_{100} = 10.5$ (*stima puntuale* di $1/p$), $\alpha = 0.04$. Poiché $\alpha/2 = 0.02$, utilizzando R si ha:

```
> alpha<-1-0.96
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 2.053749
> zalpha<-qnorm(1-alpha/2,mean=0,sd=1)
> n<-100
> medcamp<-10.5
>
> a2<-n*medcamp^2
> a1<- -(2*n*medcamp-zalpha^2)
> a0<-n-zalpha^2
> polyroot(c(a0,a1,a2))
[1] 0.07644102+0i 0.11365260-0i
```

L'intervallo di confidenza approssimato di grado $1 - \alpha = 0.96$ per p è dunque $(0.0764, 0.1137)$, ossia è bassa la probabilità che l'ago sia imperfetto per l'intera popolazione. Si nota inoltre che la stima puntuale di p, ossia $\widehat{p} = 1/\overline{x}_{100} = 0.095$ è compresa nell'intervallo. ◊

⋆ Metodo alternativo

Possiamo calcolare esplicitamente le radici dell'equazione di secondo grado in p

$$\underline{c}_n, \overline{c}_n = \frac{2n\overline{x}_n - z_{\alpha/2}^2 \pm z_{\alpha/2}\sqrt{z_{\alpha/2}^2 + 4n\overline{x}_n(\overline{x}_n - 1)}}{2n\overline{x}_n^2}$$

Se nelle stime del limite inferiore e superiore dell'intervallo di confidenza per p trascuriamo i termini che tendono a zero più rapidamente di $1/\sqrt{n}$ si ha:

$$\frac{2n\overline{x}_n - z_{\alpha/2}^2 \pm z_{\alpha/2}\sqrt{z_{\alpha/2}^2 + 4n\overline{x}_n(\overline{x}_n - 1)}}{2n\overline{x}_n^2} \simeq \frac{1}{\overline{x}_n} \pm \frac{z_{\alpha/2}}{\overline{x}_n^2}\sqrt{\frac{\overline{x}_n(\overline{x}_n - 1)}{n}}.$$

Si giunge così alla seguente proposizione:

Proposizione 12.2 *Sia $(x_1, x_2, \ldots, x_n)$ un campione osservato di ampiezza n estratto da una popolazione geometrica modificata di parametro p. Se la dimensione del campione è elevata, una stima approssimata dell'intervallo di confidenza di grado $1-\alpha$ per p è*

$$\frac{1}{\overline{x}_n} - \frac{z_{\alpha/2}}{\overline{x}_n^2}\sqrt{\frac{\overline{x}_n(\overline{x}_n-1)}{n}} < p < \frac{1}{\overline{x}_n} + \frac{z_{\alpha/2}}{\overline{x}_n^2}\sqrt{\frac{\overline{x}_n(\overline{x}_n-1)}{n}} \tag{12.7}$$

dove $\overline{x}_n$ denota la media campionaria.

Si nota che l'intervallo di confidenza per p è centrato in $1/\overline{x}_n$, in accordo con il valore medio $E(X) = 1/p$.

Relativamente all'Esempio 12.5, dalla Proposizione 12.2 si ottiene l'approssimazione:

```
> (1/medcamp)-(zalpha/medcamp^2)*sqrt(medcamp*(medcamp-1)/n)
[1] 0.07663329
> (1/medcamp)+(zalpha/medcamp^2)*sqrt(medcamp*(medcamp-1)/n)
[1] 0.1138429
```

► (**Intervallo di confidenza per il parametro λ di una popolazione di Poisson**)

Consideriamo una popolazione di Poisson descritta da una variabile aleatoria $X \sim \mathcal{P}(\lambda)$ con funzione di probabilità

$$p_X(x) = \frac{\lambda^x}{x!}\, e^{-\lambda}, \qquad x = 0, 1, \ldots \qquad (\lambda > 0).$$

Il valore medio di una variabile aleatoria di Poisson è $E(X) = \lambda$ e la varianza è $\mathrm{Var}(X) = \lambda$ ed entrambi dipendono dal parametro non noto λ. Ricaviamo che

$$E(\overline{X}_n) = \lambda, \qquad \mathrm{Var}(\overline{X}_n) = \frac{\lambda}{n}.$$

Applicando il teorema centrale di convergenza si ha che la variabile aleatoria

$$\frac{\overline{X}_n - \lambda}{\sqrt{\lambda/n}} = \sqrt{n}\,\frac{\overline{X}_n - \lambda}{\sqrt{\lambda}}$$

converge in distribuzione ad una variabile aleatoria normale standard. Per campioni sufficientemente numerosi, l'intervallo di confidenza di grado $1-\alpha$ per il parametro λ può essere determinato richiedendo che

$$P\Big(-z_{\alpha/2} < \frac{\overline{X}_n - \lambda}{\sqrt{\lambda/n}} < z_{\alpha/2}\Big) \simeq 1-\alpha. \tag{12.8}$$

La disuguaglianza

$$-z_{\alpha/2} < \frac{\overline{x}_n - \lambda}{\sqrt{\lambda/n}} < z_{\alpha/2}$$

è equivalente a

$$\Big[\sqrt{\frac{n}{\lambda}}\,(\overline{x}_n - \lambda)\Big]^2 < z^2_{\alpha/2},$$

che conduce alla disuguaglianza di secondo grado in λ

$$n\lambda^2 - \lambda(2n\overline{x}_n + z^2_{\alpha/2}) + n\overline{x}^2_n < 0. \tag{12.9}$$

Essendo il coefficiente di λ^2 positivo, le soluzioni della disuguaglianza (12.9) sono interne all'intervallo delle radici della relativa equazione di secondo grado, ossia $\underline{c}_n < \lambda < \overline{c}_n$.
Se si denota con

$$a_2 = n, \qquad a_1 = -(2n\overline{x}_n + z^2_{\alpha/2}), \qquad a_0 = n\overline{x}^2_n,$$

le radici dell'equazione $a_2\lambda^2 + a_1\lambda + a_0 = 0$ possono essere calcolate utilizzando $\mathsf{polyroot}(\mathsf{c}(\mathsf{a}_0, \mathsf{a}_1, \mathsf{a}_2))$.

Esempio 12.6 Si supponga che il numero $N(t)$ di chiamate che arrivano ad un centralino telefonico nell'intervallo $(0, t)$ sia distribuito secondo Poisson, ossia

$$P\Big(N(t) = x\Big) = \frac{(\lambda\, t)^x}{x!}\; e^{-\,\lambda\, t} \qquad (x = 0, 1, \ldots),$$

con valore medio $E[N(t)] = \lambda\, t$ e varianza $\mathrm{Var}[N(t)] = \lambda\, t$. Se in 100 osservazioni effettuate in intervalli di tempo di $t = 10$ minuti si riscontra che in media sono state effettuate 4 chiamate, si determini una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per il parametro λ.

Nel nostro caso $n = 100$, $t = 10$, $\overline{x}_{100} = 4$ (*stima puntuale* di 10λ), $\alpha = 0.05$. Poiché $\alpha/2 = 0.025$ e $1 - \alpha/2 = 0.975$, utilizzando R si ha:

```
> alpha<-1-0.95
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.959964
> zalpha<-qnorm(1-alpha/2,mean=0,sd=1)
> n<-100
> medcamp<-4
> tempo<-10
>
> a2<-n
> a1<- -(2*n*medcamp+zalpha^2)
> a0<-n*medcamp^2
> polyroot(c(a0,a1,a2))/tempo
[1] 0.3626744+0i 0.4411670-0i
```

Si nota che $z_{\alpha/2} = z_{0.025} = 1.96$. L'intervallo di confidenza approssimato di grado $1 - \alpha = 0.95$ per il parametro λ è quindi $(0.3627, 0.4412)$. La stima puntuale di λ, ossia $4/10 = 0.4$, è compresa nell'intervallo. ◊

⋆ Metodo alternativo

Possiamo calcolare esplicitamente le radici dell'equazione di secondo grado in λ:

$$\underline{c}_n, \overline{c}_n = \frac{2n\overline{x}_n + z^2_{\alpha/2} \pm z_{\alpha/2}\sqrt{z^2_{\alpha/2} + 4n\overline{x}_n}}{2n}$$

Se nelle stime del limite inferiore e superiore dell'intervallo di confidenza per λ trascuriamo i termini che tendono a zero più rapidamente di $1/\sqrt{n}$ si ha

$$\overline{x}_n + \frac{z^2_{\alpha/2}}{2n} \pm \frac{z_{\alpha/2}}{2}\sqrt{\frac{z^2_{\alpha/2}}{n^2} + \frac{4\overline{x}_n}{n}} \simeq \overline{x}_n \pm z_{\alpha/2}\sqrt{\frac{\overline{x}_n}{n}}.$$

Si giunge così alla seguente proposizione:

Proposizione 12.3 *Sia $(x_1, x_2, \ldots, x_n)$ un campione osservato di ampiezza n estratto da una popolazione di Poisson di parametro λ. Se la dimensione del campione è elevata, una stima approssimata dell'intervallo di confidenza di grado $1-\alpha$ per λ è*

$$\overline{x}_n - z_{\alpha/2}\sqrt{\frac{\overline{x}_n}{n}} < \lambda < \overline{x}_n + z_{\alpha/2}\sqrt{\frac{\overline{x}_n}{n}}\,, \tag{12.10}$$

dove $\overline{x}_n$ denota la media campionaria.

Per l'Esempio 12.6, dalla Proposizione 12.3 si ottiene l'approssimazione

```
> (medcamp-zalpha*sqrt(medcamp/n))/tempo
[1] 0.3608007
> (medcamp+zalpha*sqrt(medcamp/n))/tempo
[1] 0.4391993
```

► (**Intervallo di confidenza per il parametro ϑ di una popolazione uniforme**)

Consideriamo una popolazione uniforme descritta da una variabile aleatoria $X \sim \mathcal{U}(0, \vartheta)$ con funzione di densità di probabilità

$$f_X(x) = \frac{1}{\vartheta}, \qquad 0 < x < \vartheta.$$

Il valore medio di una variabile aleatoria uniforme è $E(X) = \vartheta/2$ e la varianza è $\mathrm{Var}(X) = \vartheta^2/12$ ed entrambi dipendono dal parametro non noto ϑ. Ricaviamo che

$$E(\overline{X}_n) = \frac{\vartheta}{2}, \qquad \mathrm{Var}(\overline{X}_n) = \frac{\vartheta^2}{12\,n}.$$

Applicando il teorema centrale di convergenza si ha che la variabile aleatoria

$$\frac{\overline{X}_n - \vartheta/2}{\vartheta/(\sqrt{12\,n})} = \sqrt{3\,n}\left(\frac{2\overline{X}_n}{\vartheta} - 1\right)$$

converge in distribuzione ad una variabile aleatoria normale standard. Per campioni sufficientemente numerosi l'intervallo di confidenza di grado $1-\alpha$ per il parametro ϑ può essere determinato richiedendo che

$$P\Big(-z_{\alpha/2} < \sqrt{3\,n}\,\Big(\frac{2\overline{X}_n}{\vartheta} - 1\Big) < z_{\alpha/2}\Big) \simeq 1-\alpha, \tag{12.11}$$

ossia

$$P\left\{2\overline{X}_n\Big(1+\frac{z_{\alpha/2}}{\sqrt{3\,n}}\Big)^{-1} < \vartheta < 2\overline{X}_n\Big(1-\frac{z_{\alpha/2}}{\sqrt{3\,n}}\Big)^{-1}\right\} \simeq 1-\alpha.$$

Sussiste quindi la seguente proposizione.

Proposizione 12.4 *Sia $(x_1, x_2, \ldots, x_n)$ un campione osservato di ampiezza n estratto da una popolazione uniforme nell'intervallo $(0, \vartheta)$. Se la dimensione del campione è elevata, una stima approssimata dell'intervallo di confidenza di grado $1-\alpha$ per ϑ è*

$$2\,\overline{x}_n\Big(1+\frac{z_{\alpha/2}}{\sqrt{3\,n}}\Big)^{-1} < \vartheta < 2\,\overline{x}_n\Big(1-\frac{z_{\alpha/2}}{\sqrt{3\,n}}\Big)^{-1}, \tag{12.12}$$

dove $\overline{x}_n$ denota la media campionaria.

Esempio 12.7 Supponiamo di considerare i tempi misurati in ore, e supposti uniformi in un intervallo $(0, \vartheta)$, necessari per soddisfare le richieste di 100 utenti che accedono ad un centro di calcolo. Se si riscontra che il tempo medio per soddisfare le richieste è di 1.5 ore, determinare una stima dell'intervallo di confidenza di grado $1-\alpha = 0.98$ per ϑ.

In questo caso $n = 100$, $\overline{x}_{100} = 1.5$ (*stima puntuale* di $\vartheta/2$ con il metodo dei momenti) e $\alpha = 0.02$. Segue che $\alpha/2 = 0.01$, $1-\alpha/2 = 0.99$. Utilizzando R si ha:

```
> alpha<-1-0.98
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 2.326348
> n<-100
> m<-1.5
>
> 2*m/(1+qnorm(1-alpha/2,mean=0,sd=1)/sqrt(3*n))
[1] 2.644776
> 2*m/(1-qnorm(1-alpha/2,mean=0,sd=1)/sqrt(3*n))
[1] 3.465451
```

Segue che $z_{\alpha/2} = z_{0.01} = 2.33$ e quindi l'intervallo di confidenza approssimato di grado $1-\alpha = 0.98$ per il parametro $\vartheta/2$ è $(1.322, 1.733)$. Si nota che la media campionaria del tempi per soddisfare le richieste degli utenti è inclusa nell'intervallo. ◊

► (**Intervallo di confidenza per il valore medio λ di una popolazione esponenziale**)

Consideriamo una popolazione esponenziale descritta da una variabile aleatoria $X \sim \mathcal{E}(\lambda)$ con funzione di densità di probabilità

$$f_X(x) = \lambda\, e^{-\lambda\, x}, \qquad x > 0 \qquad (\lambda > 0).$$

Il valore medio di una variabile aleatoria esponenziale è $E(X) = 1/\lambda$ e la varianza è $\mathrm{Var}(X) = 1/\lambda^2$ ed entrambi dipendono dal parametro non noto λ. Ricaviamo che

$$E(\overline{X}_n) = \frac{1}{\lambda}, \qquad \mathrm{Var}(\overline{X}_n) = \frac{1}{n\,\lambda^2}.$$

Applicando il teorema centrale di convergenza si ha che la variabile aleatoria

$$\frac{\overline{X}_n - 1/\lambda}{1/(\lambda\sqrt{n})} = \sqrt{n}\,\frac{\overline{X}_n - 1/\lambda}{1/\lambda} = \sqrt{n}\,(\lambda\overline{X}_n - 1)$$

converge in distribuzione ad una variabile aleatoria normale standard. Per campioni sufficientemente numerosi l'intervallo di confidenza di grado $1-\alpha$ per il parametro $1/\lambda$ può essere determinato richiedendo che

$$P\Big(-z_{\alpha/2} < \sqrt{n}\,(\lambda\overline{X}_n - 1) < z_{\alpha/2}\Big) \simeq 1-\alpha, \tag{12.13}$$

ossia

$$P\left\{\overline{X}_n\left(1+\frac{z_{\alpha/2}}{\sqrt{n}}\right)^{-1} < \frac{1}{\lambda} < \overline{X}_n\left(1-\frac{z_{\alpha/2}}{\sqrt{n}}\right)^{-1}\right\} \simeq 1-\alpha.$$

Sussiste quindi la seguente proposizione.

Proposizione 12.5 *Sia $(x_1, x_2, \ldots, x_n)$ un campione osservato di ampiezza n estratto da una popolazione esponenziale di parametro λ. Se la dimensione del campione è elevata, una stima approssimata dell'intervallo di confidenza di grado $1-\alpha$ per $1/\lambda$ è*

$$\overline{x}_n\left(1+\frac{z_{\alpha/2}}{\sqrt{n}}\right)^{-1} < \frac{1}{\lambda} < \overline{x}_n\left(1-\frac{z_{\alpha/2}}{\sqrt{n}}\right)^{-1}, \tag{12.14}$$

dove $\overline{x}_n$ denota la media campionaria.

Esempio 12.8 Si supponga che la durata delle conversazioni effettuate ad un telefono pubblico sia distribuita esponenzialmente con valore medio non noto $1/\lambda$. Se in 100 osservazioni si riscontra che in media la durata delle conversazioni degli utenti è di 3 minuti, determinare una stima dell'intervallo di confidenza di grado $1-\alpha = 0.94$ per la durata media delle conversazioni.

In questo caso $n = 100$, $\overline{x}_{100} = 3$ (*stima puntuale* di $1/\lambda$) e $\alpha = 0.06$. Segue che $\alpha/2 = 0.03$, $1-\alpha/2 = 0.97$. Utilizzando R si ha:

```
> alpha<-1-0.94
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.880794
```

```
> n<-100
> m<-3
>
> m/(1+qnorm(1-alpha/2,mean=0,sd=1)/sqrt(n))
[1] 2.525084
> m/(1-qnorm(1-alpha/2,mean=0,sd=1)/sqrt(n))
[1] 3.694942
```

Segue che $z_{\alpha/2} = z_{0.03} = 1.88$ e quindi l'intervallo di confidenza approssimato di grado $1 - \alpha = 0.94$ per il parametro $1/\lambda$ è (2.525, 3.694). Si nota che la durata media delle conversazioni dei 100 utenti è contenuta nell'intervallo. $\Diamond$

12.2 Confronto tra due popolazioni

Spesso i ricercatori sono interessati a stimare la differenza tra le medie di due distinte popolazioni. In questo caso occorre considerare due campioni casuali indipendenti $X_1, X_2, \ldots, X_{n_1}$ e $Y_1, Y_2, \ldots, Y_{n_2}$ di ampiezza n_1 e n_2 estratti rispettivamente dalle due popolazioni considerate e calcolare la media campionaria per ciascun campione, ossia $\overline{X}_{n_1}$ e $\overline{Y}_{n_2}$. Si può successivamente determinare la differenza tra le due medie campionarie $\overline{X}_{n_1} - \overline{Y}_{n_2}$. Tuttavia non si può essere certi che la differenza tra le medie campionarie $\overline{X}_{n_1} - \overline{Y}_{n_2}$ corrisponda alla differenza effettiva tra le medie delle due popolazioni. Si preferisce quindi costruire un intervallo di confidenza per la differenza tra le due medie con un certo grado di fiducia $1 - \alpha$, scelto dal decisore.

Considerate due popolazioni, descritte dalle variabili aleatorie X e Y indipendenti aventi valori medi $E(X) = \mu_1$ e $E(Y) = \mu_2$ finiti e varianze $\mathrm{Var}(X) = \sigma_1^2$ e $\mathrm{Var}(Y) = \sigma_2^2$ finite, la distribuzione della differenza $X - Y$ avrà valore medio e varianza:

$$E(X-Y) = E(X)-E(Y) = \mu_1-\mu_2, \quad \mathrm{Var}(X-Y) = \mathrm{Var}(X)+\mathrm{Var}(Y) = \sigma_1^2+\sigma_2^2.$$

Un intervallo di confidenza $(\underline{C}_n, \overline{C}_n)$ per la differenza tra due medie $\mu_1 - \mu_2$ deve essere tale che

$$P(\underline{C}_n < \mu_1 - \mu_2 < \overline{C}_n) = 1 - \alpha,$$

dove $\underline{C}_n$ e $\overline{C}_n$ sono due statistiche dipendenti dai campioni estratti dalle due popolazioni. In ricerche di questo genere è necessario conoscere il tipo di distribuzione delle due popolazioni (normali, di Bernoulli, di Poisson, ...).

L'intervallo di confidenza stimato $(\underline{c}_n, \overline{c}_n)$ può essere così interpretato:

- se il limite inferiore e il limite superiore sono entrambi negativi allora $\mu_1 - \mu_2 < 0$; ciò implica che la media della prima popolazione è inferiore alla media della seconda popolazione con un grado di confidenza $1 - \alpha$;
- se il limite inferiore e il limite superiore sono entrambi positivi allora $\mu_1 - \mu_2 > 0$; ciò implica che la media della prima popolazione è superiore alla media della seconda popolazione con un grado di confidenza $1 - \alpha$;

- se l'intervallo contiene lo zero, ossia il limite inferiore risulta negativo e il limite superiore positivo, allora con un grado di confidenza $1-\alpha$ non si può affermare che la media di una popolazione sia superiore alla media dell'altra popolazione.

12.3 Confronto tra due popolazioni normali

Siano $X_1, X_2, \ldots, X_{n_1}$ e $Y_1, Y_2, \ldots, Y_{n_2}$ due campioni casuali indipendenti di ampiezza n_1 e n_2 estratti rispettivamente da due popolazioni normali $X \sim \mathcal{N}(\mu_1, \sigma_1)$ e $Y \sim \mathcal{N}(\mu_2, \sigma_2)$. Vogliamo analizzare i seguenti problemi:

(i) determinare un intervallo di confidenza di grado $1-\alpha$ per $\mu_1 - \mu_2$ quando entrambe le varianze σ_1^2 e σ_2^2 sono note;

(ii) determinare un intervallo di confidenza di grado $1-\alpha$ per $\mu_1 - \mu_2$ quando le varianze σ_1^2 e σ_2^2 sono non note per campioni numerosi estratti dalle due popolazioni.

▶ (**Intervallo di confidenza per $\mu_1 - \mu_2$ con σ_1^2 e σ_2^2 note**)
Denotiamo con

$$\overline{X}_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \qquad \overline{Y}_{n_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$

rispettivamente le medie campionarie delle due popolazioni normali. Poiché per ipotesi i campioni casuali $X_1, X_2, \ldots, X_{n_1}$ e $Y_1, Y_2, \ldots, Y_{n_2}$ sono indipendenti, la statistica $\overline{X}_{n_1} - \overline{Y}_{n_2}$ è distribuita normalmente con valore medio e varianza

$$E(\overline{X}_{n_1} - \overline{Y}_{n_2}) = \mu_1 - \mu_2, \qquad \mathrm{Var}(\overline{X}_{n_1} - \overline{Y}_{n_2}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

ottenute ricordando la proprietà di linearità del valore medio e le proprietà della varianza per combinazioni lineari di variabili aleatorie indipendenti.

Per determinare un intervallo di confidenza di grado $1-\alpha$ per $\mu_1 - \mu_2$ quando entrambe le varianze σ_1^2 e σ_2^2 delle due popolazioni normali sono note, consideriamo la variabile aleatoria di pivot

$$Z_n = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}.$$

Tale variabile aleatoria dipende dal campione casuale e dal parametro non noto $\mu_1 - \mu_2$ (le varianze campionarie σ_1^2 e σ_2^2 delle due popolazioni sono note) ed è caratterizzata da una *densità normale standard*. Pertanto, utilizzando il *metodo pivotale* si ha

$$P\left(-z_{\alpha/2} < \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < z_{\alpha/2}\right) = 1 - \alpha.$$

Sussiste quindi la seguente proposizione.

Proposizione 12.6 *Siano $x_1, x_2, \ldots, x_{n_1}$ e $y_1, y_2, \ldots, y_{n_2}$ due campioni osservati indipendenti di ampiezza n_1 e n_2 estratti rispettivamente da due popolazioni normali $X \sim \mathcal{N}(\mu_1, \sigma_1)$ e $Y \sim \mathcal{N}(\mu_2, \sigma_2)$ le cui varianze sono note. Una stima dell'intervallo di confidenza di grado $1-\alpha$ per la differenza tra le due medie $\mu_1 - \mu_2$ è*

$$\overline{x}_{n_1} - \overline{y}_{n_2} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \overline{x}_{n_1} - \overline{y}_{n_2} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \quad (12.15)$$

dove

$$\overline{x}_n = \frac{x_1 + x_2 + \ldots + x_{n_1}}{n_1} \qquad \overline{y}_n = \frac{y_1 + y_2 + \ldots + y_{n_2}}{n_2}.$$

denotano rispettivamente le medie campionarie delle due osservazioni.

Esempio 12.9 Osservando un campione di 150 lampadine prodotte dall'industria A si riscontra che la durata media di una lampadina è 1400 ore; invece osservando un campione di 100 lampadine prodotte dall'industria B si riscontra che la durata media di una lampadina è 1200 ore. Supponendo che i campioni casuali siano stati estratti indipendentemente da due popolazioni normali $X \sim \mathcal{N}(\mu_1, \sigma_1)$ e $Y \sim \mathcal{N}(\mu_2, \sigma_2)$ con rispettive deviazioni standard $\sigma_1 = 120$ e $\sigma_2 = 80$, determinare una stima dell'intervallo di confidenza di grado $1-\alpha = 0.99$ per la differenza tra le durate medie $\mu_1 - \mu_2$ delle lampadine prodotte dalle due industrie.

In questo caso $\overline{x}_{150} = 1400$, $\overline{y}_{100} = 1200$, $\sigma_1^2 = 14400$, $\sigma_2^2 = 6400$; inoltre, essendo $\alpha = 0.01$ e $\alpha/2 = 0.005$, il valore $z_{\alpha/2} = z_{0.005}$ può essere determinato tramite R. Infatti, facendo uso della (12.15) si ha:

```
> alpha<-1-0.99
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 2.575829
> n1<-150
> n2<-100
> m1<-1400
> m2<-1200
> sigma1<-120
> sigma2<-80
>
> m1-m2-qnorm(1-alpha/2,mean=0,sd=1)*sqrt(sigma1^2/n1+sigma2^2/n2)
[1] 167.4181
> m1-m2+qnorm(1-alpha/2,mean=0,sd=1)*sqrt(sigma1^2/n1+sigma2^2/n2)
[1] 232.5819
```

Si nota che $z_{\alpha/2} = z_{0.005} = 2.575829$. La stima dell'intervallo di confidenza di grado $1-\alpha = 0.99$ per la differenza tra le durate medie $\mu_1 - \mu_2$ delle lampadine prodotte dalle due industrie è (167.42, 232.582). Poiché il limite inferiore ed il limite superiore sono positivi, si deduce che le lampadine prodotte dall'industria A hanno una durata media superiore a quella delle lampadine prodotte dall'industria B con un grado di fiducia del 99%. ◊

► (**Intervallo di confidenza per $\mu_1 - \mu_2$ con varianze non note**)

Siano $X_1, X_2, \ldots, X_{n_1}$ e $Y_1, Y_2, \ldots, Y_{n_2}$ due campioni casuali indipendenti di ampiezza n_1 e n_2 estratti da due popolazioni normali $X \sim \mathcal{N}(\mu_1, \sigma_1)$ e $Y \sim \mathcal{N}(\mu_2, \sigma_2)$ con tutti i parametri non noti. Vogliamo determinare un intervallo di confidenza di grado $1-\alpha$ per la differenza $\mu_1 - \mu_2$ delle due popolazioni per grandi valori di n_1 e n_2. Denotiamo con $S^2_{n_1}$ e $\tilde{S}^2_{n_2}$ le varianze campionarie delle due popolazioni normali. Notiamo che essendo

$$E(S^2_{n_1}) = \sigma_1^2, \qquad \lim_{n_1 \to +\infty} \mathrm{Var}[S^2_{n_1}] = 0,$$
$$E(\tilde{S}^2_{n_2}) = \sigma_2^2, \qquad \lim_{n_2 \to +\infty} \mathrm{Var}[\tilde{S}^2_{n_2}] = 0,$$

le varianze campionarie delle due popolazioni normali sono stimatori corretti e consistenti delle varianze delle due popolazioni. Quindi, quando le ampiezze dei campioni sono grandi, applicando il *metodo pivotale in forma approssimata* si ha

$$P\left(-z_{\alpha/2} < \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\mu_1 - \mu_2)}{\sqrt{S^2_{n_1}/n_1 + \tilde{S}^2_{n_2}/n_2}} < z_{\alpha/2} \right) \simeq 1 - \alpha.$$

Sussiste quindi la seguente proposizione.

Proposizione 12.7 *Siano $x_1, x_2, \ldots, x_{n_1}$ e $y_1, y_2, \ldots, y_{n_2}$ due campioni osservati indipendenti di ampiezza n_1 e n_2 estratti rispettivamente da due popolazioni normali $X \sim \mathcal{N}(\mu_1, \sigma_1)$ e $Y \sim \mathcal{N}(\mu_2, \sigma_2)$ le cui varianze sono non note. Una stima dell'intervallo di confidenza di grado $1-\alpha$ per la differenza tra le due medie $\mu_1 - \mu_2$ è*

$$\overline{x}_{n_1} - \overline{y}_{n_2} - z_{\alpha/2}\sqrt{\frac{s^2_{n_1}}{n_1} + \frac{\tilde{s}^2_{n_2}}{n_2}} < \mu_1 - \mu_2 < \overline{x}_{n_1} - \overline{y}_{n_2} + z_{\alpha/2}\sqrt{\frac{s^2_{n_1}}{n_1} + \frac{\tilde{s}^2_{n_2}}{n_2}},$$

dove $\overline{x}_{n_1}$ e $\overline{y}_{n_2}$ denotano rispettivamente le medie campionarie delle due osservazioni e dove $s^2_{n_1}$ e $\tilde{s}^2_{n_2}$ denotano rispettivamente le varianze campionarie delle due osservazioni.

Esempio 12.10 Una ditta farmaceutica è interessata a stabilire l'efficacia di un nuovo tipo di sonnifero. Un'indagine condotta su 50 pazienti mostra che il nuovo sonnifero conduce ad un numero medio di ore di sonno per individuo di 7.82 ore con una deviazione standard campionaria di 0.24 ore; un'indagine condotta su altri 100 pazienti mostra invece che il vecchio tipo di sonnifero conduce ad un numero medio di ore di sonno per individuo di 6.75 ore con una deviazione standard campionaria di 0.30 ore. Supponendo che i campioni casuali siano stati estratti indipendentemente da due popolazioni normali $\mathcal{N}(\mu_1, \sigma_1^2)$ e $\mathcal{N}(\mu_2, \sigma_2^2)$ con varianze non note, determinare una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.99$ per la differenza $\mu_1 - \mu_2$ tra i numeri medi di ore di sonno degli individui delle due popolazioni.

In questo caso $\overline{x}_{50} = 7.82$, $\overline{y}_{100} = 6.75$, $s^2_{50} = 0.0576$, $\tilde{s}^2_{100} = 0.09$; inoltre $\alpha = 0.01$ e quindi $\alpha/2 = 0.005$. Utilizzando R si ha:

```
> alpha<-1-0.99
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 2.575829
>
> n1<-50
> n2<-100
> m1<-7.82
> m2<-6.75
> s1<-0.24
> s2<-0.30
>
> m1-m2-qnorm(1-alpha/2,mean=0,sd=1)*sqrt(s1^2/n1+s2^2/n2)
[1] 0.9533175
> m1-m2+qnorm(1-alpha/2,mean=0,sd=1)*sqrt(s1^2/n1+s2^2/n2)
[1] 1.186683
```

Si nota che $z_{\alpha/2} = z_{0.005} = 2.58$. Una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.99$ per la differenza $\mu_1 - \mu_2$ tra i numeri medi di ore di sonno delle due popolazioni è (0.953, 1.187). Poiché il limite inferiore e il limite superiore sono positivi, si può dedurre che il nuovo sonnifero è più efficace rispetto al precedente sonnifero con un grado di fiducia del 99%. ◊

Esempio 12.11 (Confronto tra due server) Un account sul server A è più costoso rispetto ad un account sul server B. Tuttavia, il server A è più veloce. Per vedere se è preferibile utilizzare il server A (più veloce ma più costoso), un manager deve analizzare la velocità di esecuzione dei programmi. Un determinato algoritmo informatico viene eseguito 60 volte sul server A e 40 volte sul server B con i seguenti risultati sulle medie e deviazioni standard campionarie dei tempi di esecuzione: $\overline{x}_{60} = 6.7\,\text{sec}$, $\overline{y}_{40} = 7.5\,\text{sec}$, $s_{60} = 0.6\,\text{sec}$, $\tilde{s}_{40} = 1.2\,\text{sec}$. Supponendo che i campioni casuali siano stati estratti indipendentemente da due popolazioni normali $\mathcal{N}(\mu_1, \sigma_1^2)$ e $\mathcal{N}(\mu_2, \sigma_2^2)$ con varianze non note, determinare una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la differenza $\mu_1 - \mu_2$ tra i tempi medi di esecuzione dei programmi dei due server. Utilizzando R si ha:

```
> alpha<-1-0.95
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.959964
>
> n1<-60
> n2<-40
> m1<-6.7
> m2<-7.5
> s1<-0.6
> s2<-1.2
>
> m1-m2-qnorm(1-alpha/2,mean=0,sd=1)*sqrt(s1^2/n1+s2^2/n2)
[1] -1.201673
> m1-m2+qnorm(1-alpha/2,mean=0,sd=1)*sqrt(s1^2/n1+s2^2/n2)
[1] -0.3983269
```

Si nota che $z_{\alpha/2} = z_{0.025} = 1.96$. Una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la differenza $\mu_1 - \mu_2$ tra i tempi medi di esecuzione

dei programmi dei due server è $(-1.202, -0.398)$. Poiché il limite inferiore e il limite superiore sono negativi, si può dedurre che la velocità di esecuzione dei programmi del server A è inferiore rispetto a quella del server B con un grado di fiducia del 95%. $\Diamond$

12.4 Confronto tra due popolazioni di Bernoulli

Consideriamo una prima popolazione di Bernoulli descritta da una variabile $X \sim \mathcal{B}(1, p_1)$ con funzione di probabilità

$$p_X(x) = p_1^x (1-p_1)^{1-x}, \qquad x = 0, 1 \qquad (0 < p_1 < 1)$$

ed una seconda popolazione di Bernoulli descritta da una variabile $Y \sim \mathcal{B}(1, p_2)$ con funzione di probabilità

$$p_Y(y) = p_2^y (1-p_2)^{1-y}, \qquad y = 0, 1 \qquad (0 < p_2 < 1)$$

e siano $X_1, X_2, \ldots, X_{n_1}$ e $Y_1, Y_2, \ldots, Y_{n_2}$ due campioni casuali indipendenti di ampiezza n_1 e n_2 estratti dalle due popolazioni di Bernoulli.

Vogliamo determinare un intervallo di confidenza di grado $1-\alpha$ per la differenza $p_1 - p_2$ tra i parametri delle due popolazioni per grandi valori di n_1 e n_2. Denotiamo con $\overline{X}_{n_1}$ e $\overline{Y}_{n_2}$ rispettivamente le medie campionarie delle due popolazioni. Dal teorema centrale di convergenza segue che la variabile aleatoria

$$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \xrightarrow{d} Z,$$

converge in distribuzione ad una variabile aleatoria normale standard. Poiché

$$E(\overline{X}_{n_1}) = p_1, \quad \lim_{n_1 \to +\infty} \mathrm{Var}(\overline{X}_{n_1}) = 0, \quad E(\overline{Y}_{n_2}) = p_2, \quad \lim_{n_2 \to +\infty} \mathrm{Var}(\overline{Y}_{n_2}) = 0,$$

ossia le medie campionarie $\overline{X}_{n_1}$ e $\overline{Y}_{n_2}$ sono stimatori corretti e consistenti di p_1 e p_2, per campioni sufficientemente numerosi l'intervallo di confidenza di grado $1-\alpha$ per la differenza $p_1 - p_2$ può essere determinato supponendo che

$$P\left(-z_{\alpha/2} < \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (p_1 - p_2)}{\sqrt{\overline{X}_{n_1}(1-\overline{X}_{n_1})/n_1 + \overline{Y}_{n_2}(1-\overline{Y}_{n_2})/n_2}} < z_{\alpha/2} \right) \simeq 1 - \alpha,$$

Sussiste quindi la seguente proposizione.

Proposizione 12.8 *Siano $x_1, x_2, \ldots, x_{n_1}$ e $y_1, y_2, \ldots, y_{n_2}$ due campioni osservati indipendenti di ampiezza n_1 e n_2 estratti rispettivamente da due popolazioni di Bernoulli di parametri p_1 e p_2. Una stima approssimata dell'intervallo di confidenza di grado $1-\alpha$ per la differenza $p_1 - p_2$ è*

$$\overline{x}_{n_1} - \overline{y}_{n_2} - z_{\alpha/2} \sqrt{\frac{\overline{x}_{n_1}(1-\overline{x}_{n_1})}{n_1} + \frac{\overline{y}_{n_2}(1-\overline{y}_{n_2})}{n_2}} < p_1 - p_2$$
$$< \overline{x}_{n_1} - \overline{y}_{n_2} + z_{\alpha/2} \sqrt{\frac{\overline{x}_{n_1}(1-\overline{x}_{n_1})}{n_1} + \frac{\overline{y}_{n_2}(1-\overline{y}_{n_2})}{n_2}},$$

dove $\overline{x}_{n_1}$ e $\overline{y}_{n_2}$ denotano rispettivamente le medie campionarie delle due osservazioni.

Esempio 12.12 Un ente di ricerca demoscopica è interessato all'opinione degli elettori di due diverse città A e B in merito ad una prossima elezione politica. Su 1000 intervistati della città A, 290 hanno dichiarato che voteranno per il partito politico X; invece su 800 intervistati della città B, 264 hanno dichiarato che voteranno per lo stesso partito politico X. Sulla base di questi dati si vuole determinare una stima dell'intervallo di confidenza di grado $1-\alpha = 0.99$ per la differenza tra le frequenze relative dei votanti per quel partito nelle due città.

Possiamo supporre che le due popolazioni siano distribuite in modo bernoulliano, con parametri p_1 per la città A e p_2 per la città B. Occorre quindi determinare una stima dell'intervallo di confidenza per $p_1 - p_2$. Osserviamo che nella città A è stato osservato un campione di ampiezza $n_1 = 1000$ intervistati e 290 hanno dichiarato che voteranno per il partito politico X e quindi:

$$\overline{x}_{1000} = \frac{x_1 + x_2 + \ldots + x_{1000}}{1000} = \frac{290}{1000} = 0.29,$$

Invece nella città B è stato osservato un campione di ampiezza $n_2 = 800$ intervistati e 264 hanno dichiarato che voteranno per lo stesso partito politico X; quindi

$$\overline{y}_{800} = \frac{y_1 + y_2 + \ldots + y_{800}}{800} = \frac{264}{800} = 0.33.$$

Inoltre $\alpha = 0.01$ e quindi $\alpha/2 = 0.005$. Utilizzando R si ha:

```
> alpha<-1-0.99
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 2.575829
>
> n1<-1000
> n2<-800
> m1<-290/1000
> m2<-264/800
> rad<-sqrt(m1*(1-m1)/n1+m2*(1-m2)/n2)
>
> m1-m2-qnorm(1-alpha/2,mean=0,sd=1)*rad
[1] -0.09656717
> m1-m2+qnorm(1-alpha/2,mean=0,sd=1)*rad
[1] 0.01656717
```

Si nota che $z_{\alpha/2} = z_{0.005} = 2.58$. Inoltre, una stima dell'intervallo di confidenza di grado $1-\alpha = 0.99$ per $p_1 - p_2$ è $(-0.0966, 0.0166)$. Poiché questo intervallo include la possibilità che $p_1 = p_2$, non è possibile concludere che le frequenze relative dei votanti per il partito X nelle due città siano differenti con un grado di fiducia del 99%. ◊

Esempio 12.13 In un sondaggio su una certa trasmissione televisiva sono stati intervistati due campioni: uno di adulti (400 individui) e uno di giovani (600 individui). I giovani che hanno espresso gradimento per la trasmissione televisiva sono stati 300, gli adulti invece sono stati 100. Si desidera determinare

l'intervallo di confidenza di grado $1-\alpha = 0.95$ e di grado $1-\alpha = 0.99$ per la differenza tra le frequenze relative degli adulti e dei giovani favorevoli alla trasmissione televisiva.

Possiamo supporre che le due popolazioni siano distribuite in modo bernoulliano, con parametri p_1 per gli adulti e p_2 per i giovani. Occorre quindi determinare una stima dell'intervallo di confidenza per $p_1 - p_2$.

Osserviamo che è stato intervistato un campione di ampiezza $n_1 = 400$ di adulti e 100 hanno espresso gradimento per la trasmissione televisiva; pertanto

$$\overline{x}_{400} = \frac{100}{400} = \frac{1}{4} = 0.25.$$

È stato anche intervistato un campione di ampiezza $n_2 = 600$ di giovani e 300 hanno espresso gradimento per la trasmissione televisiva; quindi

$$\overline{y}_{600} = \frac{300}{600} = \frac{1}{2} = 0.5.$$

Utilizzando R con $1-\alpha = 0.95$ otteniamo

```
> alpha<-1-0.95
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.959964
>
> n1<-400
> n2<-600
> m1<-100/400
> m2<-300/600
> rad<-sqrt(m1*(1-m1)/n1+m2*(1-m2)/n2)
>
> m1-m2-qnorm(1-alpha/2,mean=0,sd=1)*rad
[1] -0.3083206
> m1-m2+qnorm(1-alpha/2,mean=0,sd=1)*rad
[1] -0.1916794
```

Una stima dell'intervallo di confidenza di grado $1-\alpha = 0.95$ per $p_1 - p_2$ è $(-0.3083206, -0.1916794)$.

Se invece $1-\alpha = 0.99$ utilizzando R otteniamo

```
> alpha<-1-0.99
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 2.575829
>
> n1<-400
> n2<-600
> m1<-100/400
> m2<-300/600
> rad<-sqrt(m1*(1-m1)/n1+m2*(1-m2)/n2)
>
> m1-m2-qnorm(1-alpha/2,mean=0,sd=1)*rad
[1] -0.3266463
> m1-m2+qnorm(1-alpha/2,mean=0,sd=1)*rad
[1] -0.1733537
```

Una stima dell'intervallo di confidenza di grado $1-\alpha = 0.99$ per $p_1 - p_2$ è $(-0.3266463, -0.1733537)$. Si nota che aumentando il grado di confidenza da 0.95 a 0.99 aumenta l'ampiezza dell'intervallo di confidenza stimato. Inoltre essendo $p_1 - p_2 < 0$, è possibile concludere che, relativamente alla trasmissione televisiva oggetto dell'indagine, il gradimento degli adulti è inferiore al gradimento dei giovani con un grado di fiducia del 99%. ◊

12.5 Confronto tra due popolazioni di Poisson

Consideriamo una prima popolazione di Poisson descritta da una variabile $X \sim \mathcal{P}(\lambda_1)$ con funzione di probabilità

$$p_X(x) = \frac{(\lambda_1)^x}{x!}\, e^{-\lambda_1}, \qquad x = 0, 1, \ldots \qquad (\lambda_1 > 0)$$

ed una seconda popolazione di Poisson descritta da una variabile $Y \sim \mathcal{P}(\lambda_2)$ con funzione di probabilità

$$p_Y(x) = \frac{(\lambda_2)^x}{x!}\, e^{-\lambda_2}, \qquad x = 0, 1, \ldots \qquad (\lambda_2 > 0)$$

e siano $X_1, X_2, \ldots, X_{n_1}$ e $Y_1, Y_2, \ldots, Y_{n_2}$ due campioni casuali indipendenti di ampiezza n_1 e n_2 estratti dalle due popolazioni di Poisson.

Vogliamo determinare un intervallo di confidenza di grado $1-\alpha$ per la differenza $\lambda_1 - \lambda_2$ tra i parametri delle due popolazioni per grandi valori di n_1 e n_2. Denotiamo con $\overline{X}_{n_1}$ e $\overline{Y}_{n_2}$ rispettivamente le medie campionarie delle due popolazioni. Dal teorema centrale di convergenza segue che la variabile aleatoria

$$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\lambda_1 - \lambda_2)}{\sqrt{\lambda_1/n_1 + \lambda_2/n_2}} \xrightarrow{d} Z$$

converge in distribuzione ad una variabile aleatoria normale standard. Poiché

$$E(\overline{X}_{n_1}) = \lambda_1, \quad \lim_{n_1 \to +\infty} \mathrm{Var}(\overline{X}_{n_1}) = 0, \quad E(\overline{Y}_{n_2}) = \lambda_2, \quad \lim_{n_2 \to +\infty} \mathrm{Var}(\overline{Y}_{n_2}) = 0,$$

ossia le medie campionarie $\overline{X}_{n_1}$ e $\overline{Y}_{n_2}$ sono stimatori corretti e consistenti di λ_1 e λ_2, per campioni sufficientemente numerosi l'intervallo di confidenza di grado $1-\alpha$ per la differenza $p_1 - p_2$ può essere determinato supponendo che

$$P\left(-z_{\alpha/2} < \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\lambda_1 - \lambda_2)}{\sqrt{\overline{X}_{n_1}/n_1 + \overline{Y}_{n_2}/n_2}} < z_{\alpha/2} \right) \simeq 1 - \alpha,$$

Sussiste quindi la seguente proposizione.

Proposizione 12.9 *Siano $x_1, x_2, \ldots, x_{n_1}$ e $y_1, y_2, \ldots, y_{n_2}$ due campioni osservati indipendenti di ampiezza n_1 e n_2 estratti rispettivamente da due popolazioni di Poisson di parametri λ_1 e λ_2. Una stima approssimata dell'intervallo di confidenza di grado $1-\alpha$ per la differenza $\lambda_1 - \lambda_2$ è*

$$\overline{x}_{n_1} - \overline{y}_{n_2} - z_{\alpha/2}\sqrt{\frac{\overline{x}_{n_1}}{n_1} + \frac{\overline{y}_{n_2}}{n_2}} < \lambda_1 - \lambda_2 < \overline{x}_{n_1} - \overline{y}_{n_2} + z_{\alpha/2}\sqrt{\frac{\overline{x}_{n_1}}{n_1} + \frac{\overline{y}_{n_2}}{n_2}},$$

dove $\overline{x}_{n_1}$ *e* $\overline{y}_{n_2}$ *denotano rispettivamente le medie campionarie delle due osservazioni.*

Esempio 12.14 Due incroci stradali A e B sono analizzati in base al numero di incidenti per un fissato numero di giorni. Si registrano il numero di incidenti nell'incrocio A per 50 giorni distinti e nell'incrocio B per 40 giorni distinti. Supponendo che il numero di incidenti all'incrocio A sia descritto da una variabile aleatoria di Poisson $X \sim \mathcal{P}(\lambda_1)$ e il numero di incidenti all'incrocio B sia descritto da una variabile aleatoria di Poisson $Y \sim \mathcal{P}(\lambda_2)$ si desidera determinare l'intervallo di confidenza per $\lambda_1 - \lambda_2$ di grado $1 - \alpha = 0.99$.

```
> camppoisA<-c(4, 5, 8, 0, 3, 1, 8, 2, 3, 0, 1, 2, 0, 1, 3, 1, 3,
    4, 2, 1,
+  5, 2, 0, 0, 1, 1, 3, 3, 1, 4, 5, 1, 3, 5, 0, 1, 1, 1, 4, 2,
+  6, 3, 1, 0, 2, 5, 1, 5, 1, 4)
>  length(camppoisA)
[1] 50
>
>  camppoisB<-c(1, 5, 2, 3, 2, 0, 3, 3, 0, 2, 5, 8, 1, 3, 1, 4, 2,
    6, 4, 2, 3, 2,
+  7, 5, 1, 3, 3, 4, 1, 4, 3, 3, 3, 2, 0, 2, 3, 7, 2, 1)
>  length(camppoisB)
[1] 40
```

Le frequenze assolute degli incidenti nei due incroci sono:

```
>  table(camppoisA) # frequenze assolute  incidenti in incrocio A
camppoisA
 0  1  2  3  4  5  6  8
 7 15  6  8  5  6  1  2
>  mean(camppoisA)
[1] 2.46
>
 >  table(camppoisB) # frequenze assolute  incidenti incrocio B
camppoisB
 0  1  2  3  4  5  6  7  8
 3  6  9 11  4  3  1  2  1
>  mean(camppoisB)
[1] 2.9
```

Determiniamo ora l'intervallo di confidenza per $\lambda_1 - \lambda_2$ di grado $1 - \alpha = 0.99$.

```
>  alpha<-1-0.99
>  qnorm(1-alpha/2,mean=0,sd=1)
[1] 2.575829
>  n1<- length(camppoisA)
>  n2<- length(camppoisB)
>  m1<-mean(camppoisA)
>  m2<-mean(camppoisB)
>  rad<-sqrt(m1/n1+m2/n2)
>  m1-m2- qnorm(1-alpha/2,mean=0,sd=1)*rad
[1] -1.338592
>  m1-m2+ qnorm(1-alpha/2,mean=0,sd=1)*rad
[1] 0.4585916
```

Una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.99$ per $\lambda_1 - \lambda_2$ è $(-1.3386, 0.4586)$. Poiché questo intervallo include la possibilità che $\lambda_1 = \lambda_2$, non si può concludere che il numero medio di incidenti nei due incroci siano differenti con un grado di fiducia del 99%. $\Diamond$

Occorre sottolineare che il procedimento descritto per costruire intervalli di confidenza approssimati per le differenze tra i valori medi di popolazioni di Bernoulli e di Poisson, può essere esteso anche ad altri tipi di popolazioni (geometrica, uniforme, esponenziale, ...).

Capitolo 13

Verifica delle ipotesi con R

13.1 Introduzione

Le aree più importanti dell'inferenza statistica sono la *stima dei parametri* e la *verifica delle ipotesi*. La verifica delle ipotesi interviene spesso nelle ricerche di mercato, nelle indagini sperimentali e industriali, nei sondaggi di opinione, nelle indagini sulle condizioni sociali degli abitanti di una città o di una nazione. Interviene, ad esempio, quando

- si desidera determinare se un nuovo metodo di costruzione di lampadine aumenta la durata delle stesse;
- si deve decidere se un nuovo prodotto farmaceutico è più efficace nel trattamento di una certa infezione rispetto ad un altro prodotto in commercio;
- occorre controllare se l'utilizzazione di un nuovo tipo di fertilizzante permette di aumentare la produzione annua di una certa coltura.

La verifica delle ipotesi è anche utilizzata

- per verificare se un sistema informatico non è stato infettato;
- se un aggiornamento hardware è stato efficace;
- se il numero medio di utenti simultanei nella rete è aumentato:
- se la velocità media di connessione è quella affermata dal provider di servizi Internet.

In generale gli elementi che costituiscono il punto di partenza del procedimento di verifica delle ipotesi sono una popolazione descritta da una variabile aleatoria X caratterizzata da una funzione di probabilità o densità di probabilità $f(x;\vartheta)$, un'ipotesi su di un parametro non noto ϑ della popolazione ed un campione casuale $X_1, X_2, \ldots, X_n$ estratto dalla popolazione. Occorre in primo luogo precisare il significato di ipotesi statistica.

Definizione 13.1 *Un'ipotesi statistica è un'affermazione o una congettura sul parametro non noto ϑ. Se l'ipotesi statistica specifica completamente $f(x;\vartheta)$ è detta ipotesi semplice, altrimenti è chiamata ipotesi composita.*

Per denotare un'ipotesi statistica si utilizza il carattere **H** seguito dai due punti e successivamente dall'affermazione che specifica l'ipotesi.

Esempio 13.1 Sia $X_1, X_2, \ldots, X_n$ un campione casuale estratto da una popolazione di Bernoulli $X \sim \mathcal{B}(1,p)$ e sia p la probabilità di successo. L'ipotesi statistica $\mathbf{H} : p = 0.5$ è semplice poiché specifica completamente la funzione di probabilità (ad esempio, nel lancio di una moneta si suppone che essa sia non truccata); invece, l'ipotesi statistica $\mathbf{H} : p \neq 0.5$ è composita poiché non specifica completamente la funzione di probabilità (ad esempio, nel lancio di una moneta si suppone che essa sia truccata).

Esempio 13.2 Sia $X_1, X_2, \ldots, X_n$ un campione casuale estratto da una popolazione normale $X \sim \mathcal{N}(\mu,\sigma)$ con varianza nota σ^2. Allora, l'ipotesi statistica $\mathbf{H} : \mu = 1400$ è semplice poiché, essendo nota la varianza, specifica completamente la densità; invece, l'ipotesi $\mathbf{H} : \mu \leq 1400$ è composita poiché non specifica completamente la densità. Se invece la varianza della popolazione normale non è nota, l'ipotesi statistica $\mathbf{H} : \mu = 1400$ diventa composita poiché, essendo σ^2 non nota, essa non specifica completamente la densità.

L'ipotesi soggetta a verifica è denotata con $\mathbf{H}_0$ ed è chiamata *ipotesi nulla*. Si chiama *test di ipotesi* il procedimento o regola con cui si decide, sulla base dei dati del campione, se accettare o rifiutare $\mathbf{H}_0$. La costruzione del test richiede la formulazione, in contrapposizione all'ipotesi nulla, di una proposizione alternativa. Questa proposizione prende il nome di *ipotesi alternativa* ed è di solito indicata con $\mathbf{H}_1$. L'ipotesi nulla, cioè l'ipotesi soggetta a verifica, si ha quando $\vartheta \in \Theta_0$ e l'ipotesi alternativa si ha quando $\vartheta \in \Theta_1$ e si scrive

$$\mathbf{H}_0 : \vartheta \in \Theta_0, \qquad \mathbf{H}_1 : \vartheta \in \Theta_1,$$

avendo denotato con Θ_0 e Θ_1 due sottoinsiemi disgiunti dello spazio Θ dei parametri.

Il problema della verifica delle ipotesi consiste nel determinare un test ψ che permetta di suddividere, mediante opportuni criteri, l'insieme dei possibili campioni, ossia l'insieme delle n–ple $(x_1, x_2, \ldots, x_n)$ assumibili dal vettore aleatorio $X_1, X_2, \ldots, X_n$, in due sottoinsiemi: una regione di accettazione A dell'ipotesi nulla ed una regione di rifiuto R dell'ipotesi nulla. Il test ψ può allora essere così formulato: accettare come valida l'ipotesi nulla se il campione osservato $(x_1, x_2, \ldots, x_n) \in A$ e rifiutare l'ipotesi nulla se $(x_1, x_2, \ldots, x_n) \in R$. Nel caso si verifichi che l'ipotesi nulla sia falsa, l'ipotesi alternativa sarà vera e viceversa. Spesso si usa dire che l'ipotesi nulla $\mathbf{H}_0$ deve essere verificata in alternativa all'ipotesi $\mathbf{H}_1$.

Nel seguire questo tipo di ragionamento si può incorrere in due tipi di errori:

- *rifiutare l'ipotesi nulla* $\mathbf{H}_0$ *nel caso in cui tale ipotesi sia vera*; si dice allora che si commette un errore di tipo *I* e si denota la probabilità di commettere tale errore con

$$\alpha(\vartheta) = P(\text{rifiutare } \mathbf{H}_0|\vartheta), \qquad \vartheta \in \Theta_0;$$

- *accettare l'ipotesi nulla* $\mathbf{H}_0$ *nel caso in cui tale ipotesi sia falsa*; si dice allora che si commette un errore di tipo *II* e si denota la probabilità di commettere tale errore con

$$\beta(\vartheta) = P(\text{accettare } \mathbf{H}_0|\vartheta), \qquad \vartheta \in \Theta_1.$$

Un concetto importante è quello di *misura della regione critica*.

Definizione 13.2 *Sia* ψ *un test per verificare l'ipotesi nulla* $\mathbf{H}_0 : \vartheta \in \Theta_0$ *in alternativa all'ipotesi* $\mathbf{H}_1 : \vartheta \in \Theta_1$. *Si definisce misura della regione critica del test* ψ *(o livello di significatività del test* ψ*) la seguente probabilità*

$$\alpha = \sup_{\vartheta \in \Theta_0} \alpha(\vartheta).$$

La misura della regione critica (livello di significatività α) di un test fornisce quindi la probabilità massima di commettere un errore del *I* tipo al variare di $\vartheta \in \Theta_0$, ossia la *probabilità massima di rifiutare l'ipotesi nulla quando essa è vera*. Ciò è riassunto in Tabella 13.1.

Tabella 13.1: Errori di tipo I e II

	Rifiutare $\mathbf{H}_0$	Accettare $\mathbf{H}_0$
$\mathbf{H}_0$ vera	Errore del I tipo Probabilità α	Decisione esatta Probabilità $1-\alpha$
$\mathbf{H}_0$ falsa	Decisione esatta Probabilità $1-\beta$	Errore del II tipo Probabilità β

Esiste un'analogia in ambito giudiziario che può chiarire i concetti precedenti. In tribunale una persona sottoposta ad un processo viene ritenuta innocente fino alla sentenza definitiva. L'ipotesi nulla è quindi "l'imputato è innocente"; invece, l'ipotesi alternativa è "l'imputato è colpevole". L'errore di tipo I consiste nel condannare un innocente, mentre l'errore di tipo II consiste nell'assolvere un colpevole. Riassumiamo questi concetti nella Tabella 13.2.

In generale per campioni casuali di fissata ampiezza, se si diminuisce la probabilità di commettere un errore di tipo *I* aumenta la probabilità di commettere un errore di tipo *II* e viceversa. Nella costruzione del test conviene quindi fissare la probabilità di commettere un errore di tipo *I* e cercare un test ψ che minimizzi la probabilità di commettere un errore di tipo *II*. La giustificazione del fissare

Tabella 13.2: Errori di tipo I e II in ambito giudiziario

Decisione statistica dopo il test	Imputato condannato	Imputato assolto
$\mathbf{H}_0$ vera: l'imputato è innocente	Errore del I tipo Probabilità α	Decisione esatta Probabilità $1-\alpha$
$\mathbf{H}_0$ falsa: l'imputato è colpevole	Decisione esatta Probabilità $1-\beta$	Errore del II tipo Probabilità β

la probabilità di commettere un errore di *I* tipo (che solitamente si sceglie piccola) deriva dal fatto che di solito le ipotesi vengono formulate in maniera tale che l'errore di tipo *I* sia più grave e quindi il decisore desidera imporre che la probabilità di commettere tale errore sia piccola. Ad esempio, nell'ambito giudiziario scegliere come ipotesi nulla "l'imputato è innocente" significa ritenere che condannare un innocente sia un errore più grave che assolvere un colpevole.

Solitamente la probabilità di commettere un errore di tipo *I* si sceglie uguale a 0.05, 0.01, 0.001 ed il test viene rispettivamente detto *statisticamente significativo*, *statisticamente molto significativo* e *statisticamente estremamente significativo*. Infatti, quanto minore è il valore di α tanto maggiore è la credibilità di un eventuale rifiuto dell'ipotesi nulla.

I test statistici sono di due tipi:

- *test bilaterali* (detti anche *test bidirezionali*);

- *test unilaterali* (detti anche *test unidirezionali*).

Un test bilaterale è il seguente

$$\mathbf{H}_0 : \vartheta = \vartheta_0$$
$$\mathbf{H}_1 : \vartheta \neq \vartheta_0,$$

mentre il *test unilaterale sinistro* e *test unilaterale destro* sono rispettivamente i seguenti

$$\mathbf{H}_0 : \vartheta \leq \vartheta_0 \qquad \mathbf{H}_0 : \vartheta \geq \vartheta_0$$
$$\mathbf{H}_1 : \vartheta > \vartheta_0 \qquad \mathbf{H}_1 : \vartheta < \vartheta_0,$$

avendo fissato a priori un *livello di significatività* α.

Le conclusioni dei test statistici unilaterali e bilaterali dipendono dal livello di significatività α, scelto a priori dal decisore per verificare l'ipotesi nulla $\mathbf{H}_0$.

Spesso, nei test statistici si calcola anche il *livello di significatività osservato*, noto come *p-value*. Il *p-value* si basa su una statistica del test $\widehat{\xi}_n$, che dipende dal campione osservato e dal test statistico considerato.

Il *p-value* è definito come la probabilità, supposta vera l'ipotesi $\mathbf{H}_0$, che la statistica del test $\widehat{\xi}_n$ assuma un valore uguale o più estremo di quello effettivamente osservato ξ_{os}. Essendo una probabilità il *p-value* è un numero compreso tra 0 e 1. Calcolando il *p-value* è possibile comportarsi come segue:

Criterio del p-value

- se $p > \alpha$, l'ipotesi $\mathbf{H}_0$ non può essere rifiutata;

- se $p \leq \alpha$, l'ipotesi $\mathbf{H}_0$ deve essere rifiutata.

Tuttavia, nel seguire questo ragionamento occorre prestare molta attenzione quando il valore di p è vicino ad α. Nel seguito vedremo come calcolare il *p-value* per i vari test statistici unilaterali e bilaterali utilizzando R.

Storicamente, nel 1925 Ronald Aylmer Fisher (1890-1962) sviluppò la teoria alla base del *p-value* (*p-value approach*) e nel 1933 Jerzy Neyman (1894-1981) e Egon Pearson (1895-1980) svilupparono la teoria dei test di ipotesi statistiche (*fixed alpha approach*).

Nel condurre un test statistico è importante fissare il livello di significatività α prima di calcolare il *p-value*. Se si calcola prima il *p-value*, il decisore potrebbe scegliere il livello di significatività α in funzione del risultato desiderato in modo da accettare o rigettare l'ipotesi nulla $\mathbf{H}_0$.

Nel seguito analizzeremo alcuni test unilaterali e bilaterali per popolazioni normali e successivamente indicheremo come procedere per costruire test statistici per i valori medi di altri tipi di popolazioni utilizzando il teorema centrale di convergenza.

13.2 Popolazione normale

Utilizzando test bilaterali e unilaterali, desideriamo affrontare i seguenti problemi:

(i) Verifica di ipotesi sul valore medio μ nel caso in cui la varianza σ^2 della popolazione normale è nota;

(ii) Verifica di ipotesi sul valore medio μ nel caso in cui la varianza della popolazione normale è non nota;

(iii) Verifica di ipotesi sulla varianza σ^2 nel caso in cui il valore medio μ della popolazione normale è noto;

(iv) Verifica di ipotesi sulla varianza σ^2 nel caso in cui il valore medio della popolazione normale è non noto.

13.2.1 Test su μ con varianza σ^2 nota

Sia $X_1, X_2, \ldots, X_n$ un campione casuale estratto da una popolazione normale descritta da una variabile aleatoria $X \sim \mathcal{N}(\mu, \sigma)$ con varianza nota σ^2.
$\Rightarrow$ **Test bilaterale:** Si considerino le ipotesi:

$$\mathbf{H}_0 : \mu = \mu_0, \qquad \mathbf{H}_1 : \mu \neq \mu_0$$

Essendo la varianza nota, l'ipotesi $\mathbf{H}_0$ è semplice, mentre l'ipotesi $\mathbf{H}_1$ è composita. Quando $\mathbf{H}_0$ è vera, in analogia a quanto visto per gli intervalli di confidenza, gioca un ruolo fondamentale la variabile aleatoria (statistica del test)

$$Z_n = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}},$$

che è distribuita secondo una normale standard. Occorre osservare che Z_n è una statistica (non è una variabile di pivot) poiché dipende esclusivamente dal campione casuale essendo μ_0 e σ^2 noti. Il test bilaterale ψ di misura α per le ipotesi considerate è il seguente:

- si accetti $\mathbf{H_0}$ se $\quad -z_{\alpha/2} < \dfrac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}$

- si rifiuti $\mathbf{H_0}$ se $\quad \dfrac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2} \quad$ oppure $\quad \dfrac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2}$

Nella Figura 13.1 è rappresentata la densità normale standard e sono sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla per il test bilaterale. Il valore $z_{\alpha/2}$ è calcolato tramite $\mathsf{qnorm}(1-\alpha/2, \mathsf{mean}=0, \mathsf{sd}=1)$.

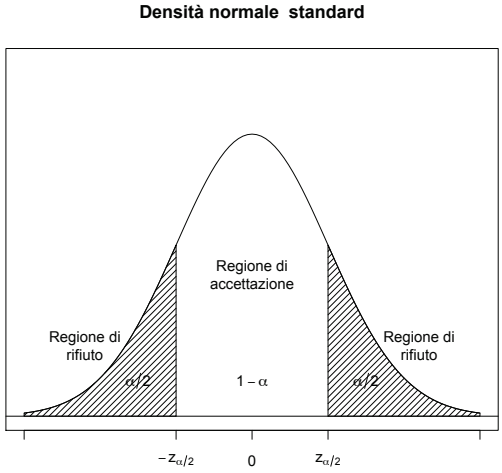

Figura 13.1: Densità normale standard e regioni di accettazione e di rifiuto dell'ipotesi nulla per il test bilaterale

La Figura 13.1 è ottenuta con il seguente codice:

```
>curve(dnorm(x,mean=0,sd=1),from=-3, to=3,axes=FALSE,ylim=c(0,0.5),
+xlab="",ylab="",main="Densita' normale standard")
>text(0,0.05,expression(1-alpha))
>text(0,0.2,"Regione di\naccettazione")
```

```
>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(-2.2,0.1,"Regione di\nrifiuto")
>text(1.5,0.05,expression(alpha/2))
>text(2.2,0.1,"Regione di\nrifiuto")
>box()
```

Denotando con

$$z_{os} = \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}},$$

la stima osservata della statistica del test, desideriamo calcolare il *p-value* per il test bilaterale considerato (livello di significatività osservato):

$$\begin{aligned} pvalue &= P(Z_n < -|z_{os}|) + P(Z_n > |z_{os}|) = 2\,P(Z_n > |z_{os}|) \\ &= 2\Big[1 - P(Z_n \leq |z_{os}|)\Big], \end{aligned}$$

e corrisponde alla probabilità, supposta vera l'ipotesi nulla $\mathbf{H}_0 : \mu = \mu_0$, che la statistica del test Z_n assuma un valore uguale o più estremo di quello effettivamente osservato z_{os}.

In R il *p-value* del test bilaterale considerato può essere così calcolato:

$$2 * (1 - \mathsf{pnorm}(\mathsf{abs}(\mathsf{z_{os}}), \mathsf{mean} = 0, \mathsf{sd} = 1))$$

Si nota che il *p-value* non dipende dalla scelta del livello di significatività α.

Esempio 13.3 Una ditta produttrice di lampadine sostiene che la durata media di un certo tipo di lampadine prodotte sia $\mu = 1600$ ore, con una deviazione standard $\sigma = 120$ ore. Viene analizzato un campione di 100 lampadine e si riscontra una durata media di 1570 ore. Si desidera costruire il test di misura $\alpha = 0.05$ per verificare l'ipotesi nulla $\mathbf{H}_0 : \mu = 1600$ in alternativa all'ipotesi $\mathbf{H}_1 : \mu \neq 1600$.
Occorre applicare un test di verifica di ipotesi bilaterale. Nel nostro caso $\alpha = 0.05$, $\mu_0 = 1600$, $\sigma = 120$, $n = 100$, $\overline{x}_{100} = 1570$. Utilizzando R, risulta:

```
> alpha<-0.05
> mu0<-1600
> sigma<-120
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.959964
> n<-100
> meancamp<-1570
> (meancamp-mu0)/(sigma/sqrt(n))
```

```
[1] -2.5
>
> pvalue<-2*(1-pnorm(2.5,mean=0,sd=1))
> pvalue
[1] 0.01241933
```

Si nota che $z_{\alpha/2} = 1.959964$ e $z_{os} = -2.5$ cade al di fuori della regione di accettazione; occorre quindi rifiutare l'ipotesi nulla con un livello di significatività del 5%. Si nota anche che $pvalue < \alpha$ e quindi anche il criterio del *p-value* consiglia di rifiutare l'ipotesi nulla.

$\Rightarrow$ **Test unilaterale sinistro:** Si considerino le ipotesi:

$$\mathbf{H}_0 : \mu \leq \mu_0, \qquad \mathbf{H}_1 : \mu > \mu_0$$

Le ipotesi $\mathbf{H}_0$ e $\mathbf{H}_1$ sono entrambe composite. Scegliamo il più grande valore di μ tale che l'potesi $\mathbf{H}_0$ sia vera. Il test unilaterale sinistro ψ di misura α per le ipotesi considerate è il seguente:

- si accetti $\mathbf{H_0}$ se $\dfrac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < z_\alpha$

- si rifiuti $\mathbf{H_0}$ se $\dfrac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$

Nella Figura 13.2 è rappresentata la densità normale e sono sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla del test unilaterale sinistro. Il valore z_α è calcolato tramite $\mathsf{qnorm}(1-\alpha, \mathsf{mean}=0, \mathsf{sd}=1)$.
La Figura 13.2 è ottenuta con il seguente codice:

```
> curve(dnorm(x,mean=0,sd=1),from=-3, to=3,axes=FALSE,ylim=c(0,0.5)
    ,xlab="",
+ ylab="",main="Densita' normale standard")
> text(0,0.05,expression(1-alpha))
> text(0,0.2,"Regione di\naccettazione")
> axis(1,c(-3,-1,0,1,3),c("","  "," ",expression(z[alpha]),""))
> 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))
> text(2.2,0.1,"Regione di\nrifiuto")
> box()
```

Calcoliamo ora *p-value* per il test unilaterale sinistro considerato:

$$pvalue = P(Z_n > z_{os}) = 1 - P(Z_n \leq z_{os}),$$

dove $z_{os} = (\overline{x}_n - \mu_0)/(\sigma/\sqrt{n})$ è la stima della statistica del test. Il *p-value* corrisponde alla probabilità, supposta vera l'ipotesi nulla $\mathbf{H}_0 : \mu \leq \mu_0$, che la statistica del test Z_n assuma un valore uguale o più estremo di quello effettivamente osservato z_{os}.

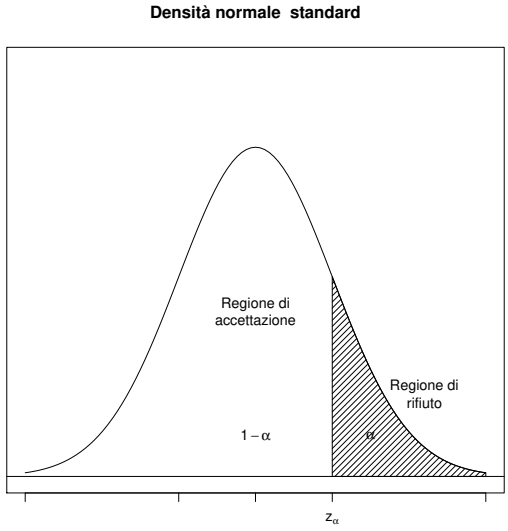

Figura 13.2: Densità normale standard e regioni di accettazione e di rifiuto dell'ipotesi nulla per il test unilaterale sinistro

In R il *p-value* può essere così calcolato:

$$1 - \mathsf{pnorm}(\mathsf{z_{os}}, \mathsf{mean} = 0, \mathsf{sd} = 1)$$

⇒ **Test unilaterale destro:** Si considerano le ipotesi:

$$\mathbf{H}_0 : \mu \geq \mu_0, \qquad \mathbf{H}_1 : \mu < \mu_0 \tag{13.1}$$

Le ipotesi $\mathbf{H_0}$ e $\mathbf{H_1}$ sono entrambe composite. Scegliamo il più piccolo valore di μ tale che l'potesi $\mathbf{H}_0$ è vera. Il test unilaterale destro ψ di misura α per le ipotesi considerate è il seguente

- si accetti $\mathbf{H_0}$ se $\dfrac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} > -z_\alpha$

- si rifiuti $\mathbf{H_0}$ se $\dfrac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$

Nella Figura 13.3 è rappresentata la densità normale standard e sono sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla del test unilaterale destro. Il valore $-z_\alpha$ è calcolato tramite $\mathsf{qnorm}(\alpha, \mathsf{mean} = 0, \mathsf{sd} = 1)$.
La Figura 13.3 è ottenuta con il seguente codice:

```
> curve(dnorm(x,mean=0,sd=1),from=-3, to=3,axes=FALSE,ylim=c(0,0.5)
    ,xlab="",
```

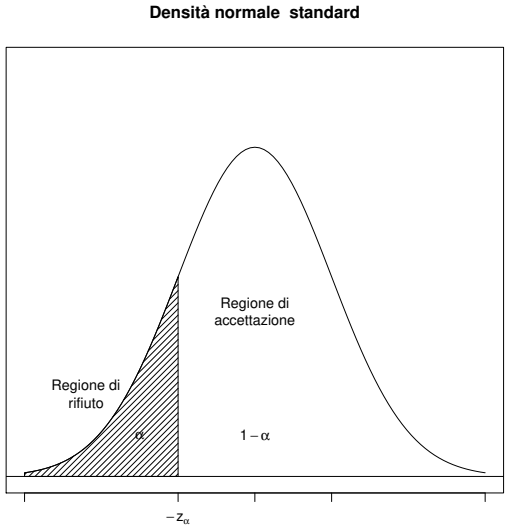

Figura 13.3: Densità normale standard e regioni di accettazione e di rifiuto dell'ipotesi nulla per il test unilaterale destro

```
+ ylab="",main="Densita' normale standard")
> text(0,0.05,expression(1-alpha))
> text(0,0.2,"Regione di\naccettazione")
> axis(1,c(-3,-1,0,1,3),c("",expression(-z[alpha])," "," ",""))
> 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)
> abline(h=0)
> text(-1.5,0.05,expression(alpha))
> text(-2.2,0.1,"Regione di\nrifiuto")
> box()
```

Calcoliamo ora *p-value* per il test unilaterale destro considerato:

$$pvalue = P(Z_n \leq z_{os}),$$

dove $z_{os} = (\overline{x}_n - \mu_0)/(\sigma/\sqrt{n})$ è la stima della statistica del test. Il *p-value* corrisponde alla probabilità, supposta vera l'ipotesi nulla $\mathbf{H}_0 : \mu \geq \mu_0$, che la statistica del test Z_n assuma un valore uguale o più estremo di quello effettivamente osservato z_{os}.

In R il *p-value* può essere così calcolato:

$$\mathsf{pnorm}(\mathsf{z_{os}}, \mathsf{mean} = 0, \mathsf{sd} = 1)$$

Esempio 13.4 Un'industria produttrice di un nuovo tipo di fertilizzante assicura che l'utilizzazione di tale prodotto per la produzione di una certa coltura con-

durrà ad una produzione media annua maggiore o uguale a $1800\,kg$ per ettaro, con una deviazione standard di $120\,kg$. Un'azienda agricola desidera controllare se l'utilizzazione di questo nuovo tipo di fertilizzante permetta effettivamente di ottenere la produzione media annua dichiarata dall'industria. Per risolvere il problema l'azienda osserva il raccolto ottenuto in 60 differenti appezzamenti di un ettaro ciascuno ed ottiene una produzione media $\overline{x}_{60} = 1780\,kg$. Si desidera costruire il test di misura $\alpha = 0.05$ per verificare l'ipotesi nulla $\mathbf{H}_0 : \mu \geq 1800$ in alternativa all'ipotesi $\mathbf{H}_1 : \mu < 1800$.
Occorre applicare un test di verifica di ipotesi unilaterale destro. Nel nostro caso $\alpha = 0.05$, $\mu_0 = 1800$, $n = 60$, $\overline{x}_{60} = 1780$, $\sigma = 120$. Utilizzando R, si ha

```
> alpha<-0.05
> mu0<-1800
> sigma<-120
> qnorm(alpha,mean=0,sd=1)
[1] -1.644854
> n<-60
> meancamp<-1780
> (meancamp-mu0)/(sigma/sqrt(n))
[1] -1.290994
>
> pvalue<-pnorm(-1.290994,mean=0,sd=1)
> pvalue
[1] 0.09835288
```

Si nota che $-z_\alpha = -1.644854$ e $z_{os} = -1.290994$ cade nella regione di accettazione. Occorre quindi accettare l'ipotesi nulla con un livello di significatività del 5%. Si nota anche che $pvalue > \alpha$ e quindi anche il criterio del *p-value* consiglia di accettare l'ipotesi nulla.

13.2.2 Test su μ con varianza non nota

Sia $X_1, X_2, \ldots, X_n$ un campione casuale estratto da una popolazione normale con varianza non nota σ^2.
$\Rightarrow$ **Test bilaterale:** Si considerino le ipotesi:

$$\mathbf{H}_0 : \mu = \mu_0 \qquad \mathbf{H}_1 : \mu \neq \mu_0$$

Essendo la varianza non nota, entrambe le ipotesi sono composite. Quando $\mathbf{H}_0$ è vera, in analogia a quanto visto per gli intervalli di confidenza, gioca un ruolo fondamentale la variabile aleatoria (statistica del test)

$$T_n = \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}}\,.$$

che è distribuita con legge di Student con $n-1$ gradi di libertà. Notiamo che T_n è una statistica poiche' dipende esclusivamente dal campione casuale considerato. Il test bilaterale ψ di misura α per le ipotesi considerate è il seguente:

- si accetti $\mathbf{H_0}$ se $\quad -t_{\alpha/2,n-1} < \dfrac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} < t_{\alpha/2,n-1}$

- si rifiuti $\mathbf{H_0}$ se $\quad \dfrac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} < -t_{\alpha/2,n-1} \quad$ oppure $\quad \dfrac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} > t_{\alpha/2,n-1}$

Nella Figura 13.4 è rappresentata la densità di Student con $n-1$ gradi di libertà e sono sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla per il test bilaterale. Il valore $t_{\alpha/2}$ è calcolato tramite $\mathsf{qt}(1-\alpha/2, \mathsf{df} = \mathsf{n}-1)$.

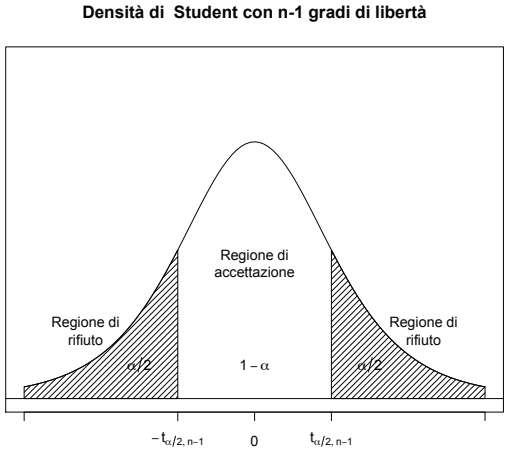

Figura 13.4: Densità di Student con $n-1$ gradi di libertà e regioni di accettazione e di rifiuto dell'ipotesi nulla per il test bilaterale

La Figura 13.4 è ottenuta con il seguente codice:

```
> curve(dt(x,df=5),from=-3, to=3,axes=FALSE,ylim=c(0,0.5),xlab="",
+ ylab="",main="Densita' di Student con n-1 gradi di liberta' ")
> text(0,0.05,expression(1-alpha))
> text(0,0.2,"Regione di\naccettazione")
> 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)
> 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(-2.2,0.1,"Regione di\nrifiuto")
> text(1.5,0.05,expression(alpha/2))
> text(2.2,0.1,"Regione di\nrifiuto")
> box()
```

Denotando con

$$t_{os} = \frac{\overline{x}_n - \mu_0}{(s_n/\sqrt{n})},$$

la stima della statistica del test, calcoliamo ora il *p-value* per il test bilaterale considerato (livello di significatività osservato):

$$\begin{aligned} pvalue &= P(T_n < -|t_{os}|) + P(T_n > |t_{os}|) = 2\,P(T_n > |t_{os}|) \\ &= 2\Big[1 - P(T_n \leq |t_{os}|)\Big], \end{aligned}$$

In R il *p-value* può essere così calcolato:

$$2 * (1 - \mathsf{pt}(\mathsf{abs}(\mathsf{t_{os}}), \mathsf{df} = \mathsf{n} - 1))$$

Esempio 13.5 Una ditta dichiara che un certo tipo di tubi hanno un contenuto medio di rame del $23\,gr$. La ditta desidera controllare se la quantità di rame presente nei tubi prodotti è quella richiesta. A tal fine, analizza un campione di 20 tubi e riscontra un contenuto medio di rame di $\overline{x}_{20} = 23.5\,gr$ con una deviazione standard campionaria di $s = 0.24\,gr$. Si desidera utilizzare il test di misura $\alpha = 0.01$ per verificare l'ipotesi nulla $\mathbf{H}_0 : \mu = 23$ in alternativa all'ipotesi $\mathbf{H}_1 : \mu \neq 23$.
Occorre applicare un test di verifica di ipotesi bilaterale. Nel nostro caso $\alpha = 0.01$, $\mu_0 = 23$, $n = 20$, $\overline{x}_{20} = 23.5$, $s = 0.24$. Utilizzando R, risulta:

```
> alpha<-0.01
> mu0<-23
> n<-20
> qt(1-alpha/2,df=n-1)
[1] 2.860935
> meancamp<-23.5
> devcamp<-0.24
> (meancamp-mu0)/(devcamp/sqrt(n))
[1] 9.31695
>
> pvalue<-2*(1-pt(9.31695,df=n-1))
> pvalue
[1] 1.624559e-08
```

Si nota che $t_{\alpha/2,n-1} = 2.860935$ e $t_{os} = 9.31695$ cade al di fuori della regione di accettazione. Occorre quindi rifiutare l'ipotesi nulla con un livello di significatività dell'1%. La ditta ne deduce che i tubi prodotti non hanno mantenuto la proporzione richiesta di rame. Essendo $pvalue < \alpha$ anche il criterio del *p-value* consiglia di rifiutare l'ipotesi nulla.

Esempio 13.6 Una compagnia aerea afferma che il peso medio del bagaglio dei passeggeri dei suoi voli di linea è $19.8\,kg$. La compagnia desidera sottoporre a verifica tale ipotesi con un livello di significatività dell'1%. A tal fine, considera un campione di 100 passeggeri e riscontra un peso medio campionario di $20.2\,kg$ con una deviazione standard campionaria di $3.6\,kg$.

Occorre utilizzare un test bilaterale $\mathbf{H}_0 : \mu = 19.8$ in alternativa all'ipotesi $\mathbf{H}_1 : \mu \neq 19.8$. Nel caso considerato $\mu_0 = 19.8$, $\alpha = 0.01$, $n = 100$, $\overline{x}_{100} = 20.2$ e $s_{100} = 3.6$. Utilizzando R, si ha:

```
> alpha<-0.01
> mu0<-19.8
> n<-100
> qt(1-alpha/2,df=n-1)
[1] 2.626405
> meancamp<-20.2
> devcamp<-3.6
> (meancamp-mu0)/(devcamp/sqrt(n))
[1] 1.111111
>
> pvalue<-2*(1-pt(1.111111,df=n-1))
> pvalue
[1] 0.2692118
```

Si nota che $t_{\alpha/2,n-1} = 2.63$ e $t_{os} = 1.11$ cade nella regione di accettazione. L'ipotesi nulla $\mathbf{H}_0$ deve essere accettata con il livello di significatività richiesto. Essendo $pvalue > \alpha$ anche il criterio del *p-value* consiglia di accettare l'ipotesi nulla.

$\Rightarrow$ **Test unilaterale sinistro:** Si considerano le ipotesi

$$\mathbf{H}_0 : \mu \leq \mu_0 \qquad \mathbf{H}_1 : \mu > \mu_0$$

Entrambe le ipotesi $\mathbf{H}_0$ e $\mathbf{H}_1$ sono composite. Il test unilaterale sinistro ψ di misura α per le ipotesi considerate è il seguente:

- si accetti $\mathbf{H_0}$ se $\dfrac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} < t_{\alpha,n-1}$

- si rifiuti $\mathbf{H_0}$ se $\dfrac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} > t_{\alpha,n-1}$

Nella Figura 13.5 è rappresentata la densità di Student con $n-1$ gradi di libertà e sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla per il test unilaterale sinistro. Il valore $t_{\alpha,n-1}$ è calcolato tramite $\mathsf{qt}(1-\alpha, \mathsf{df} = \mathsf{n}-1)$. La Figura 13.5 è ottenuta con il seguente codice:

```
> curve(dt(x,df=5),from=-3, to=3,axes=FALSE,ylim=c(0,0.5),xlab="",
+ ylab="",main="Densita' di  Student con n-1 gradi di liberta' ")
> text(0,0.05,expression(1-alpha))
> text(0,0.2,"Regione di\naccettazione")
> axis(1,c(-3,-1,0,1,3),c("",expression(-t[list(alpha,n-1)]),"  ","
    ",""))
> vals<-seq(-3,-1,length=100)
> x<-c(-3,vals,-1,-3)
> 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))
> text(-2.2,0.1,"Regione di\nrifiuto")
> box()
```

Calcoliamo ora *p-value* per il test unilaterale sinistro:

$$pvalue = P(T_n > t_{os}) = 1 - P(T_n \leq t_{os}),$$

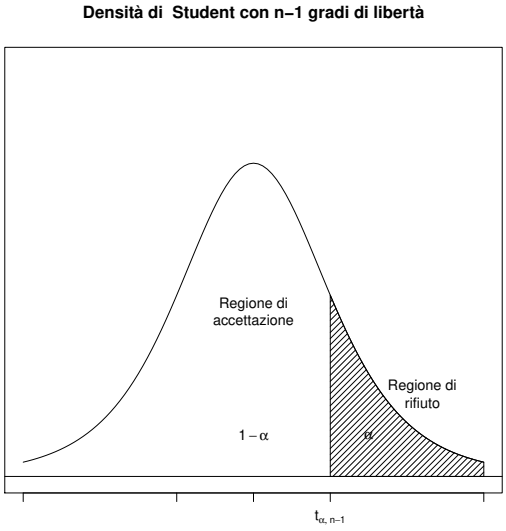

Figura 13.5: Densità di Student con $n-1$ gradi di libertà e regioni di accettazione e di rifiuto dell'ipotesi nulla per il test unilaterale sinistro

dove $t_{os} = (\overline{x}_n - \mu_0)/(s_n/\sqrt{n})$ è la stima della statistica del test considerato. In R il *p-value* può essere così calcolato:

$$1 - \mathsf{pt}(\mathsf{t_{os}}, \mathsf{df} = \mathsf{n} - 1)$$

$\Rightarrow$ **Test unilaterale destro:** Si considerino le ipotesi:

$$\mathbf{H}_0 : \mu \geq \mu_0 \qquad \mathbf{H}_1 : \mu < \mu_0$$

Il test unilaterale destro ψ di misura α per le ipotesi considerate è il seguente:

- si accetti $\mathbf{H_0}$ se $\dfrac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} > -t_{\alpha,n-1}$

- si rifiuti $\mathbf{H_0}$ se $\dfrac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} < -t_{\alpha,n-1}$

Nella Figura 13.6 è rappresentata la densità di Student con $n-1$ gradi di libertà e sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla del test unilaterale destro. Il valore $-t_{\alpha,n-1}$ è calcolato tramite $\mathsf{qt}(\alpha, \mathsf{df} = \mathsf{n} - 1)$. La Figura 13.6 è ottenuta con il seguente codice:

```
> curve(dt(x,df=5),from=-3, to=3,axes=FALSE,ylim=c(0,0.5),xlab="",
+ ylab="",main="Densita' di  Student con n-1 gradi di liberta' ")
> text(0,0.05,expression(1-alpha))
```

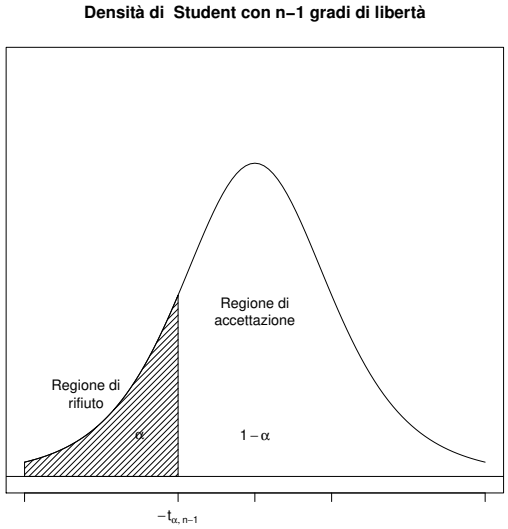

Figura 13.6: Densità di Student con $n-1$ gradi di libertà e regioni di accettazione e di rifiuto dell'ipotesi nulla per il test unilaterale destro

```
> text(0,0.2,"Regione di\naccettazione")
> axis(1,c(-3,-1,0,1,3),c("","","",expression(t[list(alpha,n-1)])
    ,""))
> 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))
> text(2.2,0.1,"Regione di\nrifiuto")
> box()
```

Calcoliamo ora *p-value* per il test unilaterale destro:

$$pvalue = P(T_n \leq t_{os}),$$

dove $t_{os} = (\overline{x}_n - \mu_0)/(s_n/\sqrt{n})$ è la stima della statistica del test considerato. In R il *p-value* può essere così calcolato:

$$\mathsf{pt}(\mathsf{t_{os}}, \mathsf{df} = \mathsf{n} - 1)$$

Esempio 13.7 Il reddito medio annuale di una famiglia che abita in una fissata provincia non supera 12500 Euro. Si desidera sottoporre a verifica tale ipotesi con un livello di significatività dell'1%. A tal fine, si considera un campione di 80 famiglie e si riscontra che il reddito medio campionario è 12000 Euro con una deviazione standard campionaria di 1500 Euro.

Occorre applicare un test di verifica di ipotesi unilaterale sinistro $\mathbf{H}_0 : \mu \leq 12500$ in alternativa all'ipotesi $\mathbf{H}_1 : \mu > 12500$. Nel nostro caso $\alpha = 0.01$, $\mu_0 = 12500$, $n = 80$, $\overline{x}_{80} = 12000$, $s_{80} = 1500$. Utilizzando R, si ha

```
> alpha<-0.01
> mu0<-12500
> n<-80
> qt(1-alpha,df=n-1)
[1] 2.374482
> meancamp<-12000
> devcamp<-1500
> (meancamp-mu0)/(devcamp/sqrt(n))
[1] -2.981424
>
> pvalue<-1-pt(-2.981424,df=n-1)
> pvalue
[1] 0.9980935
```

Si nota che $t_{\alpha,n-1} = 2.374482$ e $t_{os} = -2.981424$ cade nella regione di accettazione. Occorre quindi accettare l'ipotesi sul reddito medio annuale delle famiglie con un livello di significatività dell'1%. Essendo $pvalue > \alpha$, anche il criterio del *p-value* consiglia di accettare l'ipotesi nulla.

Esempio 13.8 Una ditta produttrice di pneumatici afferma che la durata media di un certo tipo di pneumatici è di almeno $50000\,km$. Un'officina desidera controllare se l'utilizzazione di questo tipo di pneumatici permetta effettivamente di ottenere la durata media dichiarata dalla ditta produttrice. Per risolvere il problema, sottopone a prove su strada un campione di 40 pneumatici dello stesso tipo e misura una durata media $\overline{x} = 49400\,km$ con una deviazione standard $s = 2500\,km$. Si desidera costruire il test di misura $\alpha = 0.05$ per verificare l'ipotesi nulla $\mathbf{H}_0 : \mu \geq 50000$ in alternativa all'ipotesi $\mathbf{H}_1 : \mu < 50000$.
Occorre applicare un test di verifica di ipotesi unilaterale destro. Nel nostro caso $\alpha = 0.05$, $\mu_0 = 50000$, $n = 40$, $\overline{x}_{40} = 49400$, $s_{40} = 2500$. Utilizzando R, si ha

```
> alpha<-0.05
> mu0<-50000
> n<-40
> qt(alpha,df=n-1)
[1] -1.684875
> meancamp<-49400
> devcamp<-2500
> (meancamp-mu0)/(devcamp/sqrt(n))
[1] -1.517893
>
> pvalue<-pt(-1.517893,df=n-1)
> pvalue
[1] 0.06855337
```

Si nota che $-t_{\alpha,n-1} = -1.684875$ e $t_{os} = -1.517893$ cade nella regione di accettazione. Occorre quindi accettare l'ipotesi della ditta produttrice sulla la durata media di un certo tipo di pneumatici con un livello di significatività del

5%. Essendo $pvalue > \alpha$ anche il criterio del *p-value* consiglia di accettare l'ipotesi nulla.

13.2.3 Test su σ^2 con valore medio noto

Sia $X_1, X_2, \ldots, X_n$ un campione casuale estratto da una popolazione normale con valore medio noto μ.
$\Rightarrow$ **Test bilaterale:** Si considerino le ipotesi:

$$\mathbf{H}_0 : \sigma^2 = \sigma_0^2, \qquad \mathbf{H}_1 : \sigma^2 \neq \sigma_0^2$$

Essendo il valore medio noto, l'ipotesi $\mathbf{H}_0$ è semplice; invece l'ipotesi $\mathbf{H}_1$ è composita. Quando $\mathbf{H}_0$ è vera, in analogia a quanto visto per gli intervalli di confidenza, gioca un ruolo rilevante la variabile aleatoria (statistica del test)

$$V_n = \sum_{i=1}^{n} \Big(\frac{X_i - \mu}{\sigma_0} \Big)^2 = \frac{(n-1)\, S_n^2}{\sigma_0^2} + \Big(\frac{\overline{X}_n - \mu}{\sigma_0 / \sqrt{n}} \Big)^2$$

che è distribuita con legge chi–quadrato con n gradi di libertà.
Il test bilaterale ψ di misura α per le ipotesi considerate è il seguente

- si accetti $\mathbf{H_0}$ se $\chi^2_{1-\alpha/2,n} < \sum_{i=1}^{n} \Big(\frac{x_i - \mu}{\sigma_0} \Big)^2 < \chi^2_{\alpha/2,n}$

- si rifiuti $\mathbf{H_0}$ se $\sum_{i=1}^{n} \Big(\frac{x_i - \mu}{\sigma_0} \Big)^2 < \chi^2_{1-\alpha/2,n}$ oppure $\sum_{i=1}^{n} \Big(\frac{x_i - \mu}{\sigma_0} \Big)^2 > \chi^2_{\alpha/2,n}$

Nella Figura 13.7 è rappresentata la densità chi-quadrato con n gradi di libertà e sono sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla per il test bilaterale. Il valore $\chi^2_{1-\alpha/2,n}$ si calcola con $\mathsf{qchisq}(\alpha/2, \mathsf{df} = \mathsf{n})$ e il valore $\chi^2_{\alpha/2,n}$ si calcola con $\mathsf{qchisq}(1-\alpha/2, \mathsf{df} = \mathsf{n})$.
La Figura 13.7 è ottenuto con il seguente codice:

```
> curve(dchisq(x,df=6),from=0, to=12,axes=FALSE,ylim=c(0,0.15),
+ xlab="",ylab="",main="Densita' chi-quadrato con n gradi di
    liberta'")
> text(5,0.02,expression(1-alpha))
> text(4.8,0.10,"Regione di\naccettazione")
> axis(1,c(0,3,5.5,8,12),c("",expression({chi^2}[list(1-alpha/2,n)
    ]),
+ expression(n-2),expression({chi^2}[list(alpha/2,n)]),""))
> vals<-seq(0,3,length=100)
> x<-c(0,vals,3,0)
> y<-c(0,dchisq(vals,df=6),0,0)
> polygon(x,y,density=20,angle=45)
> vals<-seq(8,12,length=100)
> x<-c(8,vals,12,7)
> y<-c(0,dchisq(vals,df=6),0,0)
> polygon(x,y,density=20,angle=45)
> abline(h=0)
```

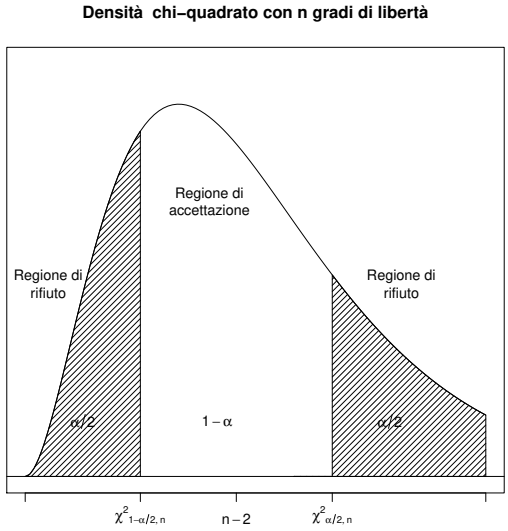

Figura 13.7: Densità chi-quadrato con n gradi di libertà e zone di accettazione e rifiuto dell'ipotesi nulla per il test bilaterale.

```
> text(1.5,0.02,expression(alpha/2))
> text(0.6,0.07,"Regione di\nrifiuto")
> text(9.5,0.02,expression(alpha/2))
> text(9.8,0.07,"Regione di\nrifiuto")
> box()
```

Esempio 13.9 Un'industria che produce batterie al litio dichiara che hanno una durata di vita media di 3 anni con una deviazione standard di 1 anno. Estratto un campione di 50 batterie, si riscontra che la media campionaria è di 3.1 anni e la deviazione standard campionaria è $\sqrt{0.9}$ anni. L'industria desidera verificare se la varianza dichiarata per le batterie prodotte sia effettivamente quella dichiarata. Si desidera costruire il test di misura $\alpha = 0.05$ per verificare l'ipotesi nulla $\mathbf{H}_0 : \sigma^2 = 1$ in alternativa all'ipotesi $\mathbf{H}_1 : \sigma^2 \neq 1$.
In questo caso $\mu = 3$, $\sigma_0 = 1$, $n = 50$, $\overline{x}_{50} = 3.1$, $s^2_{50} = 0.9$, $\alpha = 0.05$. Utilizzando R, risulta:

```
> alpha<-0.05
> mu<-3
> sigma02<-1
> n<-50
> medcamp<-3.1
> varcamp<-0.9
> qchisq(alpha/2,df=n)
[1] 32.35736
> qchisq(1-alpha/2,df=n)
```

```
[1] 71.4202
> (n-1)*varcamp/sigma02+n*(medcamp-mu)**2/sigma02
[1] 44.6
```

Si nota che $\chi^2_{1-\alpha/2,50} = 32.36$, $\chi^2_{\alpha/2,50} = 71.42$ e $\chi^2 = 44.6$. Poichè il valore osservato $\chi^2 = 44.6$ è compreso nella regione di accettazione, si accetta l'ipotesi nulla e l'industria attesta che la varianza della durata delle batterie prodotte non si discosta significativamente da 1 con un livello di significatività del 5%.

⇒ **Test unilaterale sinistro:** Si desidera verificare le ipotesi:

$$\mathbf{H}_0 : \sigma^2 \leq \sigma_0^2 \qquad \mathbf{H}_1 : \sigma^2 > \sigma_0^2$$

Entrambe le ipotesi sono composite. Il test unilaterale sinistro ψ di misura α per le ipotesi considerate è

- si accetti $\mathbf{H_0}$ se $\displaystyle\sum_{i=1}^{n}\Big(\frac{x_i-\mu}{\sigma_0}\Big)^2 < \chi^2_{\alpha,n}$

- si rifiuti $\mathbf{H_0}$ se $\displaystyle\sum_{i=1}^{n}\Big(\frac{x_i-\mu}{\sigma_0}\Big)^2 > \chi^2_{\alpha,n}$

Nella Figura 13.8 è rappresentata la densità chiquadrato con n gradi di libertà e sono sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla per il test unilaterale sinistro. ll valore $\chi^2_{\alpha,n}$ si calcola con $\mathsf{qchisq}(1-\alpha, \mathsf{df} = \mathsf{n})$.
La Figura 13.8 è ottenuta con il seguente codice:

```
> curve(dchisq(x,df=6),from=0, to=12,axes=FALSE,ylim=c(0,0.15),xlab
    ="",ylab="",
+ main="Densita' chi-quadrato con n gradi di liberta' ")
> text(4,0.02,expression(1-alpha))
> text(4,0.10,"Regione di\naccettazione")
> axis(1,c(0,2,4,6,12),c("","",expression(n-2),
+ expression({chi^2}[list(alpha,n)]),""))
> 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(8.5,0.02,expression(alpha))
> text(8.8,0.08,"Regione di\nrifiuto")
> box()
```

⇒ **Test unilaterale destro:** Si considerino le ipotesi:

$$\mathbf{H}_0 : \sigma^2 \geq \sigma_0^2 \qquad \mathbf{H}_1 : \sigma^2 < \sigma_0^2$$

Entrambe le ipotesi sono composite. Il test unilaterale destro ψ di misura α per le ipotesi considerate è

- si accetti $\mathbf{H_0}$ se $\displaystyle\sum_{i=1}^{n}\Big(\frac{x_i-\mu}{\sigma_0}\Big)^2 > \chi^2_{1-\alpha,n}$

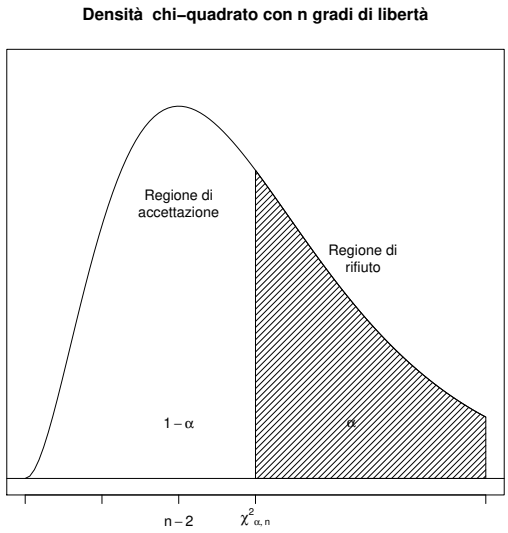

Figura 13.8: Densità chi-quadrato con n gradi di libertà e zone di accettazione e rifiuto dell'ipotesi nulla per il test unilaterale sinistro.

- si rifiuti $\mathbf{H_0}$ se $\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma_0}\right)^2 < \chi^2_{1-\alpha,n}$

Nella Figura 13.9 è rappresentata la densità chi-quadrato con n gradi di libertà e sono sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla per il test unilaterale destro. Il valore $\chi^2_{1-\alpha,n}$ si calcola con $\mathsf{qchisq}(\alpha, \mathsf{df} = \mathsf{n})$.
La Figura 13.9 è ottenuta con il seguente codice

```
> curve(dchisq(x,df=6),from=0, to=12,axes=FALSE,ylim=c(0,0.15),xlab
    ="",ylab="",
+ main="Densita' chi-quadrato con n gradi di liberta' ")
> text(4,0.02,expression(1-alpha))
> text(4,0.10,"Regione di\naccettazione")
> axis(1,c(0,2,4,6,12),c("",expression({chi^2}[list(1-alpha,n)]),
+ expression(n-2),
+ "",""))
> 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)
> text(1.2,0.02,expression(alpha))
> text(0.5,0.07,"Regione di\nrifiuto")
> abline(h=0)
> box()
```

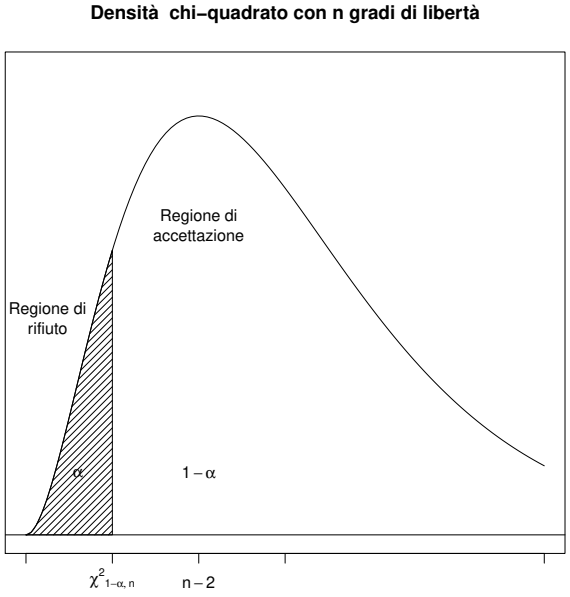

Figura 13.9: Densità chi-quadrato con n gradi di libertà e zone di accettazione e rifiuto dell'ipotesi nulla per il test unilaterale destro.

13.2.4 Test su σ^2 con valore medio non noto

Sia $X_1, X_2, \ldots, X_n$ un campione casuale estratto da una popolazione normale con valore medio noto μ.
$\Rightarrow$ **Test bilaterale:** Si considerino le ipotesi:

$$\mathbf{H}_0 : \sigma^2 = \sigma_0^2, \qquad \mathbf{H}_1 : \sigma^2 \neq \sigma_0^2$$

Entrambe le ipotesi sono composite. Quando l'ipotesi $\mathbf{H}_0$ è vera, in analogia a quanto visto per gli intervalli di confidenza, gioca un ruolo rilevante la variabile aleatoria (statistica del test)

$$Q_n = \frac{(n-1)\, S_n^2}{\sigma_0^2} = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

che è distribuita con legge chi–quadrato con $n-1$ gradi di libertà.

Il test bilaterale ψ di misura α per le ipotesi considerate è

- si accetti $\mathbf{H_0}$ se $\chi^2_{1-\alpha/2,n-1} < \dfrac{(n-1)\, s_n^2}{\sigma_0^2} < \chi^2_{\alpha/2,n-1}$

- si rifiuti $\mathbf{H_0}$ se $\dfrac{(n-1)\, s_n^2}{\sigma_0^2} < \chi^2_{1-\alpha/2,n-1}$ oppure $\dfrac{(n-1)\, s_n^2}{\sigma_0^2} > \chi^2_{\alpha/2,n-1}$

Nella Figura 13.10 è rappresentata la densità chi-quadrato con $n-1$ gradi di libertà e sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla. ll valore $\chi^2_{1-\alpha/2,n-1}$ si calcola con $\mathsf{qchisq}(\alpha/2, \mathsf{df} = \mathsf{n}-1)$ e il valore $\chi^2_{\alpha/2,n-1}$ si calcola con $\mathsf{qchisq}(1-\alpha/2, \mathsf{df} = \mathsf{n}-1)$.

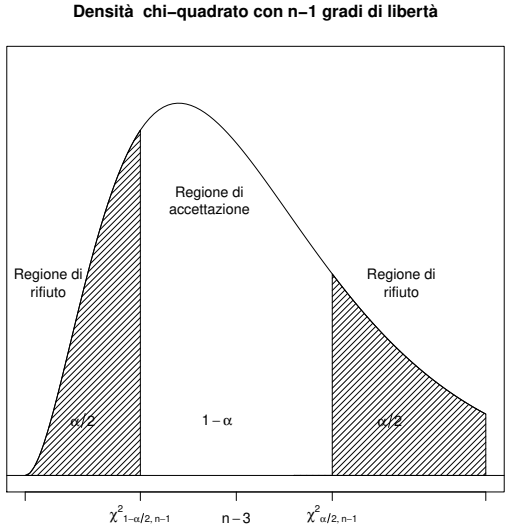

Figura 13.10: Densità chi-quadrato con $n-1$ gradi di libertà e zone di accettazione e rifiuto dell'ipotesi nulla per il test bilaterale.

Esempio 13.10 Un'industria che produce batterie al litio dichiara che la durata in anni ha una deviazione standard di 1 anno. Estratto un campione di 50 batterie, si riscontra che la deviazione standard campionaria è $\sqrt{0.9}$ anni. L'industria desidera verificare se la varianza dichiarata per le batterie prodotte sia effettivamente quella dichiarata. Si desidera costruire il test di misura $\alpha = 0.05$ per verificare l'ipotesi nulla $\mathbf{H}_0 : \sigma^2 = 1$ in alternativa all'ipotesi $\mathbf{H}_1 : \sigma^2 \neq 1$. In questo caso $\sigma_0 = 1$, $n = 50$, $s^2_{50} = 0.9$, $\alpha = 0.05$. Utilizzando R, risulta:

```
> alpha<-0.05
> sigma02<-1
> n<-50
> varcamp<-0.9
> qchisq(alpha/2,df=n-1)
[1] 31.55492
> qchisq(1-alpha/2,df=n-1)
[1] 70.22241
>  (n-1)*varcamp/sigma02
[1] 44.1
```

Si nota che $\chi^2_{1-\alpha/2,49} = 31.55$, $\chi^2_{\alpha/2,49} = 70.22$ e $\chi^2 = 44.1$. Poichè il valore osservato $\chi^2 = 44.1$ è compreso nella regione di accettazione, si accetta l'ipotesi nulla e l'industria attesta l'ipotesi sulla varianza della durata delle batterie con un livello di significatività del 5%.

⇒ **Test unilaterale sinistro:** Si considerino le ipotesi statistiche

$$\mathbf{H}_0 : \sigma^2 \leq \sigma_0^2, \qquad \mathbf{H}_1 : \sigma^2 > \sigma_0^2.$$

Il test unilaterale sinistro ψ di misura α per le ipotesi considerate è

- si accetti $\mathbf{H_0}$ se $\dfrac{(n-1)s_n^2}{\sigma_0^2} < \chi^2_{\alpha,n-1}$

- si rifiuti $\mathbf{H_0}$ se $\dfrac{(n-1)s_n^2}{\sigma_0^2} > \chi^2_{\alpha,n-1}$

Nella Figura 13.11 è rappresentata la densità chi-quadrato con $n-1$ gradi di libertà sono sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla per il test unilaterale sinistro. ll valore $\chi^2_{\alpha,n-1}$ si calcola con $\mathsf{qchisq}(1-\alpha, \mathsf{df} = \mathsf{n}-1)$.

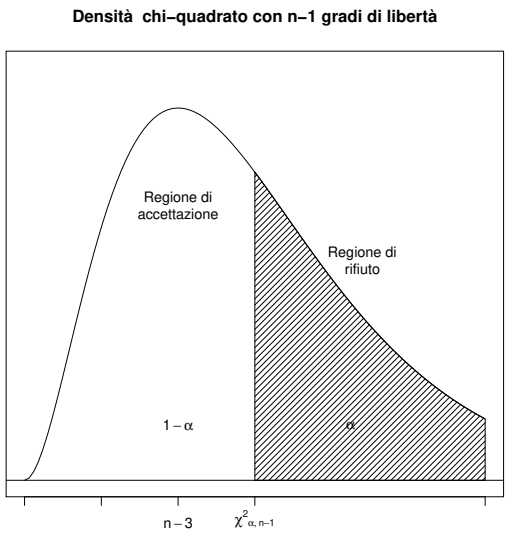

Figura 13.11: Densità chi-quadrato con $n-1$ gradi di libertà e zone di accettazione e rifiuto dell'ipotesi nulla per l'ipotesi unilaterale sinistra.

⇒ **Test unilaterale destro:** Si considerino le ipotesi statistiche:

$$\mathbf{H}_0 : \sigma^2 \geq \sigma_0^2 \qquad \mathbf{H}_1 : \sigma^2 < \sigma_0^2.$$

Il test unilaterale destro ψ di misura α per le ipotesi considerate è

- si accetti $\mathbf{H_0}$ se $\dfrac{(n-1)s_n^2}{\sigma_0^2} > \chi^2_{1-\alpha,n-1}$

- si rifiuti $\mathbf{H_0}$ se $\dfrac{(n-1)s_n^2}{\sigma_0^2} < \chi^2_{1-\alpha,n-1}$

Nella Figura 13.12 è rappresentata la densità chi-quadrato con $n-1$ gradi di libertà e sono sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla. ll valore $\chi^2_{1-\alpha,n-1}$ si calcola con $\mathsf{qchisq}(\alpha, \mathsf{df} = \mathsf{n}-1)$.

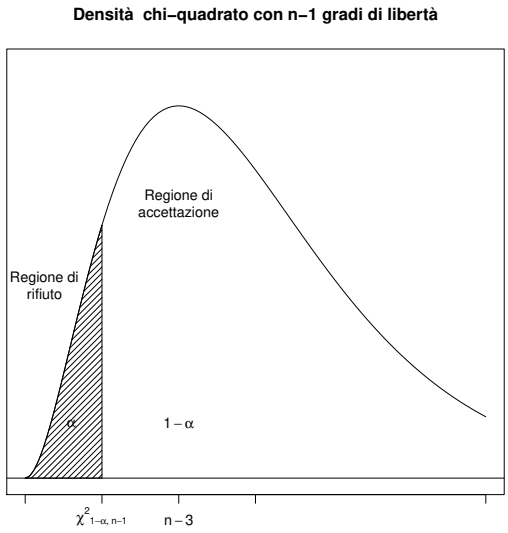

Figura 13.12: Densità chi-quadrato con $n-1$ gradi di libertà e zone di accettazione e rifiuto dell'ipotesi nulla per il test unilaterale destro.

Desideriamo infine determinare dei test statistici bilaterali e unilaterali per il valore medio di popolazioni non normali per campioni numerosi utilizzando il teorema centrale di convergenza.

13.3 Test statistici per grandi campioni

Quando l'ampiezza del campione è grande, per una popolazione descritta da una variabile aleatoria X caratterizzata da valore medio μ e varianza σ^2, entrambi finiti, si può utilizzare il teorema centrale di convergenza ricordando che la variabile aleatoria

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z,$$

converge in distribuzione ad una variabile normale standard.

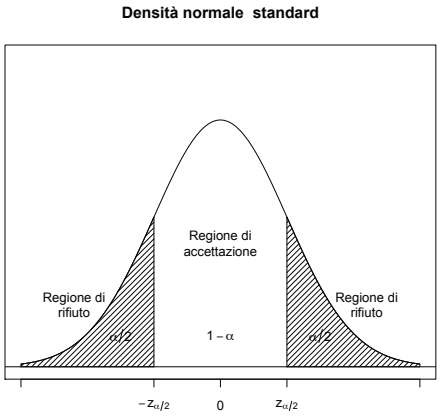

Figura 13.13: Test bilaterale

⇒ **Test bilaterale approssimato:** Per campioni numerosi, il test bilaterale ψ di misura α per le ipotesi

$$\mathbf{H}_0 : \mu = \mu_0, \qquad \mathbf{H}_1 : \mu \neq \mu_0$$

considera come variabile aleatoria

$$\frac{\overline{X}_n - \mu_0}{\sigma_0/\sqrt{n}},$$

dove σ_0 è la deviazione standard della popolazione quando $\mu = \mu_0$. Tale variabile aleatoria deve dipendere soltanto dal campione casuale e costituisce la statistica del test. Come evidenziato in Figura 13.13, il test bilaterale ψ di misura α è il seguente:

- si accetti $\mathbf{H_0}$ se $\quad -z_{\alpha/2} < \dfrac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} < z_{\alpha/2}$

- si rifiuti $\mathbf{H_0}$ se $\quad \dfrac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} < -z_{\alpha/2} \quad$ oppure $\quad \dfrac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} > z_{\alpha/2}$

dove $z_{\alpha/2}$ si calcola in R con $\mathsf{qnorm}(1 - \alpha/2, \mathsf{mean} = 0, \mathsf{sd} = 1)$.
⇒ **Test unilaterale sinistro approssimato:** Per campioni numerosi, il test unilaterale sinistro ψ di misura α per le ipotesi

$$\mathbf{H}_0 : \mu \leq \mu_0, \qquad \mathbf{H}_1 : \mu > \mu_0$$

è (vedi Figura 13.14):

- si accetti $\mathbf{H_0}$ se $\dfrac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} < z_\alpha$

- si rifiuti $\mathbf{H_0}$ se $\dfrac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} > z_\alpha$

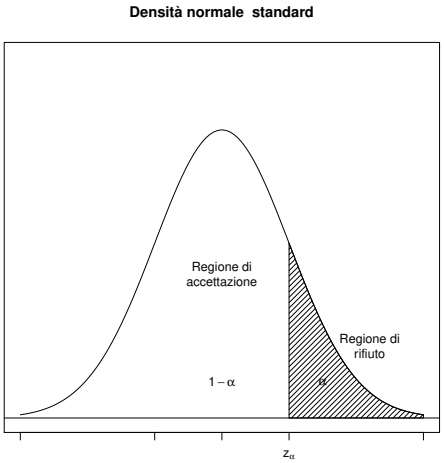

Figura 13.14: Test unilaterale sinistro

dove z_α si calcola in R con $\mathsf{qnorm}(1-\alpha, \mathsf{mean}=0, \mathsf{sd}=1)$.
$\Rightarrow$ **Test unilaterale destro approssimato:** Per campioni numerosi, il test unilaterale destro ψ di misura α per le ipotesi

$$\mathbf{H}_0 : \mu \geq \mu_0, \qquad \mathbf{H}_1 : \mu < \mu_0$$

è (vedi Figura 13.15):

- si accetti $\mathbf{H_0}$ se $\dfrac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} > -z_\alpha$

- si rifiuti $\mathbf{H_0}$ se $\dfrac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} < -z_\alpha$

dove $-z_\alpha$ si calcola in R con $\mathsf{qnorm}(\alpha, \mathsf{mean}=0, \mathsf{sd}=1)$.

13.3.1 Popolazione di Bernoulli

Consideriamo una popolazione di Bernoulli descritta dalla variabile aleatoria $X \sim \mathcal{B}(p)$. Siamo interessati a costruire dei test unilaterali e bilaterali per il valore medio $E(X) = p$. Il test bilaterale può essere così formulato:

$$\mathbf{H}_0 : p = p_0$$
$$\mathbf{H}_1 : p \neq p_0,$$

mentre il *test unilaterale sinistro* e *test unilaterale destro* sono rispettivamente i seguenti

$$\mathbf{H}_0 : p \leq p_0 \qquad \mathbf{H}_0 : p \geq p_0$$
$$\mathbf{H}_1 : p > p_0 \qquad \mathbf{H}_1 : p < p_0,$$

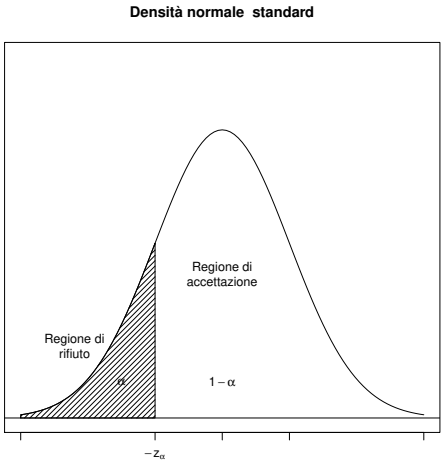

Figura 13.15: Test unilaterale destro

avendo fissato a priori un *livello di significatività* α. Essendo $\mu_0 = p_0$ e $\sigma_0^2 = p_0(1-p_0)$, nei test unilaterali e bilaterali occorre considerare

$$z_{os} = \frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} = \frac{\overline{x}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Esempio 13.11 Una ditta farmaceutica è interessata a verificare l'efficacia di un nuovo farmaco per curare una data malattia. Da un'indagine condotta su 900 pazienti affetti da questa malattia trova che il farmaco è efficace in 740 casi. Possiamo supporre che la popolazione sia distribuita secondo Bernoulli, con p che denota la probabilità che il farmaco sia efficace. Nell'Esempio 12.2 abbiamo mostrato che una stima dell'intervallo di confidenza di grado $1-\alpha = 0.95$ per p è $(0.796, 0.846)$.

Si desidera verificare l'ipotesi $\mathbf{H}_0 : p \geq 0.8$ in alternativa a $\mathbf{H}_1 : p < 0.8$ con un livello di significatività $\alpha = 0.05$. Occorre considerare un test unilaterale destro. Utilizzando R si ha:

```
> p0<-0.8
> alpha<-0.05
> qnorm(alpha,mean=0,sd=1)
[1] -1.644854
> n<-900
> meancamp<-740/900
> (meancamp-p0)/sqrt(p0*(1-p0)/n)
[1] 1.666667
```

Si nota che $-z_\alpha = -1.644854$ e $z_{os} = 1.666667$ cade nella regione di accettazione. Occorre quindi accettare l'ipotesi nulla con un livello di significatività del 5%.

13.3.2 Popolazione di Poisson

Consideriamo una popolazione di Poisson descritta dalla variabile aleatoria $X \sim \mathcal{P}(\lambda)$. Siamo interessati a costruire dei test unilaterali e bilaterali per il valore medio $E(X) = \lambda$. Il test bilaterale può essere così formulato:

$$\mathbf{H}_0 : \lambda = \lambda_0$$
$$\mathbf{H}_1 : \lambda \neq \lambda_0,$$

mentre il *test unilaterale sinistro* e *test unilaterale destro* sono rispettivamente i seguenti

$$\mathbf{H}_0 : \lambda \leq \lambda_0 \qquad \mathbf{H}_0 : \lambda \geq \lambda_0$$
$$\mathbf{H}_1 : \lambda > \lambda_0 \qquad \mathbf{H}_1 : \lambda < \lambda_0,$$

avendo fissato a priori un *livello di significatività* α. Essendo $\mu_0 = \lambda$ e $\sigma_0^2 = \lambda$, nei test unilaterali e bilaterali occorre considerare

$$z_{os} = \frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} = \frac{\overline{x}_n - \lambda_0}{\sqrt{\frac{\lambda_0}{n}}} = \sqrt{n}\,\frac{\overline{x}_n - \lambda_0}{\sqrt{\lambda_0}}$$

Esempio 13.12 Si supponga che il numero $N(t)$ di chiamate che arrivano ad un centralino telefonico nell'intervallo $(0, t)$ sia distribuito secondo Poisson con valore medio $E[N(t)] = \lambda t$. In 100 osservazioni effettuate in intervalli di tempo di $t = 10$ minuti si riscontra che in media sono state effettuate 4 chiamate. Nell'Esempio 12.6 abbiamo mostrato che una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per il parametro λ è (0.3627, 0.4412).

Ci proponiamo ora verificare l'ipotesi $\mathbf{H}_0 : 10\,\lambda \leq 3.5$ in alternativa a $\mathbf{H}_1 : 10\,\lambda > 3.5$ con un livello di significatività $\alpha = 0.05$. Occorre considerare un test unilaterale sinistro. Utilizzando R si ha:

```
> lambda0<-3.5
> alpha<-0.05
> qnorm(1-alpha,mean=0,sd=1)
[1] 1.644854
> n<-100
> meancamp<-4
> (meancamp-lambda0)/sqrt(lambda0/n)
[1] 2.672612
```

Si nota che $z_\alpha = 1.644854$ e $z_{os} = 2.672612$ cade nella regione di rifiuto. Occorre quindi rifiutare l'ipotesi nulla che $\lambda \leq 0.35$ con un livello di significatività del 5%.

13.3.3 Popolazione esponenziale

Consideriamo una popolazione esponenziale descritta dalla variabile aleatoria $X \sim \mathcal{E}(\lambda)$. Siamo interessati a costruire dei test unilaterali e bilaterali per il

valore medio $E(X) = 1/\lambda$. Il test bilaterale può essere così formulato:

$$\mathbf{H}_0 : \frac{1}{\lambda} = \frac{1}{\lambda_0}$$
$$\mathbf{H}_1 : \frac{1}{\lambda} \neq \frac{1}{\lambda_0},$$

mentre il *test unilaterale sinistro* e *test unilaterale destro* sono rispettivamente i seguenti

$$\mathbf{H}_0 : \frac{1}{\lambda} \leq \frac{1}{\lambda_0} \qquad \mathbf{H}_0 : \frac{1}{\lambda} \geq \frac{1}{\lambda_0}$$
$$\mathbf{H}_1 : \frac{1}{\lambda} > \frac{1}{\lambda_0} \qquad \mathbf{H}_1 : \frac{1}{\lambda} < \frac{1}{\lambda_0},$$

avendo fissato a priori un *livello di significatività* α. Essendo $\mu_0 = 1/\lambda_0$ e $\sigma_0^2 = 1/\lambda_0^2$, nei test unilaterali e bilaterali occorre considerare

$$z_{os} = \frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} = \frac{\overline{x}_n - \frac{1}{\lambda_0}}{\sqrt{\frac{1}{n\lambda_0^2}}} = \sqrt{n}\,(\lambda_0\,\overline{x}_n - 1)$$

Esempio 13.13 Si supponga che la durata delle conversazioni effettuate ad un telefono pubblico sia distribuita esponenzialmente con valore medio non noto $1/\lambda$. In 100 osservazioni si riscontra che in media la durata delle conversazioni degli utenti è di 3 minuti. Nell'Esempio 12.8 abbiamo mostrato che una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.94$ per il parametro $1/\lambda$ è (2.525, 3.694).

Ci proponiamo ora verificare l'ipotesi $\mathbf{H}_0 : 1/\lambda = 3.2$ in alternativa a $\mathbf{H}_1 : 1/\lambda \neq 3.2$ con un livello di significatività $\alpha = 0.06$. Utilizzando R si ha:

```
> lambda0<-1/3.2
> alpha<-0.06
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.880794
> n<-100
> meancamp<-3
> sqrt(n)*(lambda0*meancamp-1)
[1] -0.625
```

Si nota che $z_{\alpha/2} = 1.880794$ e $z_{os} = -0.625$ cade nella regione di accettazione. Occorre quindi accettare l'ipotesi nulla che $1/\lambda = 3.2$ con un livello di significatività del 6%.

Nelle Tabelle 13.3, 13.4, 13.5 riassumiamo le regioni di accettazione dei test bilaterali, unilaterali sinistri e unilaterali destri con livello di significatività α per il valore medio e della varianza di una popolazione descritta da una variabile aleatoria normale $X \sim \mathcal{N}(\mu, \sigma)$. Infine, nella Tabella 13.6 riassumiamo le regioni di accettazione dei test bilaterale, unilaterale sinistro e unilaterale destro con livello di significatività α per il valore medio della popolazione per campioni numerosi.

Tabella 13.3: Test bilaterale per una normale con significatività α

Test bilaterale	Regione di accettazione di $\mathbf{H_0}$
Test su μ con varianza σ^2 nota $\mathbf{H}_0 : \mu = \mu_0, \quad \mathbf{H}_1 : \mu \neq \mu_0$	$-z_{\alpha/2} < \dfrac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}$ $z_{\alpha/2}$ si calcola con $\mathsf{qnorm}(1 - \mathsf{alpha}/2, \mathsf{mean} = 0, \mathsf{sd} = 1)$ $pvalue = P(Z_n < -\lvert z_{os}\rvert) + P(Z_n > \lvert z_{os}\rvert) = 2\Big[1 - P(Z_n \leq \lvert z_{os}\rvert)\Big]$ il p-value si calcola con $2 * (1 - \mathsf{pnorm}(\mathsf{abs}(\mathsf{z_{os}}), \mathsf{mean} = 0, \mathsf{sd} = 1))$
Test su μ con varianza non nota $\mathbf{H}_0 : \mu = \mu_0, \quad \mathbf{H}_1 : \mu \neq \mu_0$	$-t_{\alpha/2,n-1} < \dfrac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} < t_{\alpha/2,n-1}$ $t_{\alpha/2,n-1}$ si calcola con $\mathsf{qt}(1 - \mathsf{alpha}/2, \mathsf{df} = \mathsf{n} - 1))$ $pvalue = P(T_n < -\lvert t_{os}\rvert) + P(T_n > \lvert t_{os}\rvert) = 2\Big[1 - P(T_n \leq \lvert t_{os}\rvert)\Big]$ il p-value si calcola con $2 * (1 - \mathsf{pt}(\mathsf{abs}(\mathsf{t_{os}}), \mathsf{df} = \mathsf{n} - 1))$
Test su σ^2 con valore medio μ noto $\mathbf{H}_0 : \sigma^2 = \sigma_0^2, \quad \mathbf{H}_1 : \sigma^2 \neq \sigma_0^2$	$\chi^2_{1-\alpha/2,n} < \displaystyle\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma_0}\right)^2 < \chi^2_{\alpha/2,n}$ $\chi^2_{1-\alpha/2,n}$ si calcola con $\mathsf{qchisq}(\mathsf{alpha}/2, \mathsf{df} = \mathsf{n})$ $\chi^2_{\alpha/2,n}$ si calcola con $\mathsf{qchisq}(1 - \mathsf{alpha}/2, \mathsf{df} = \mathsf{n})$
Test su σ^2 con valore medio non noto $\mathbf{H}_0 : \sigma^2 = \sigma_0^2, \quad \mathbf{H}_1 : \sigma^2 \neq \sigma_0^2$	$\chi^2_{1-\alpha/2,n-1} < \dfrac{(n-1)\, s_n^2}{\sigma_0^2} < \chi^2_{\alpha/2,n-1}$ $\chi^2_{1-\alpha/2,n-1}$ si calcola con $\mathsf{qchisq}(\mathsf{alpha}/2, \mathsf{df} = \mathsf{n} - 1)$ $\chi^2_{\alpha/2,n-1}$ si calcola con $\mathsf{qchisq}(1 - \mathsf{alpha}/2, \mathsf{df} = \mathsf{n} - 1)$

Tabella 13.4: Test unilaterale sinistro per una normale con significatività α

Test unilaterale sinistro	Regione di accettazione di $\mathbf{H_0}$
Test su μ con varianza σ^2 nota $\mathbf{H}_0 : \mu \leq \mu_0, \quad \mathbf{H}_1 : \mu > \mu_0$	$\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < z_\alpha$ z_α si calcola con $\mathsf{qnorm}(1 - \mathsf{alpha}, \mathsf{mean} = 0, \mathsf{sd} = 1)$ ..... $pvalue = P(Z_n > z_{os}) = 1 - P(Z_n \leq z_{os})$ il p-value si calcola con $1 - \mathsf{pnorm}(\mathsf{z_{os}}, \mathsf{mean} = 0, \mathsf{sd} = 1)$
Test su μ con varianza non nota $\mathbf{H}_0 : \mu \leq \mu_0, \quad \mathbf{H}_1 : \mu > \mu_0$	$\frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} < t_{\alpha,n-1}$ $t_{\alpha,n-1}$ si calcola con $\mathsf{qt}(1 - \mathsf{alpha}, \mathsf{df} = \mathsf{n} - 1))$ ..... $pvalue = P(T_n > t_{os}) = 1 - P(T_n \leq t_{os})$ il p-value si calcola con $1 - \mathsf{pt}(\mathsf{t_{os}}, \mathsf{df} = \mathsf{n} - 1)$
Test su σ^2 con valore medio μ noto $\mathbf{H}_0 : \sigma^2 \leq \sigma_0^2, \quad \mathbf{H}_1 : \sigma^2 > \sigma_0^2$	$\sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma_0} \right)^2 < \chi^2_{\alpha,n}$ $\chi^2_{\alpha,n}$ si calcola con $\mathsf{qchisq}(1 - \mathsf{alpha}, \mathsf{df} = \mathsf{n})$
Test su σ^2 con valore medio non noto $\mathbf{H}_0 : \sigma^2 \leq \sigma_0^2, \quad \mathbf{H}_1 : \sigma^2 > \sigma_0^2$	$\frac{(n-1)\, s_n^2}{\sigma_0^2} < \chi^2_{\alpha,n-1}$ $\chi^2_{\alpha,n-1}$ si calcola con $\mathsf{qchisq}(1 - \mathsf{alpha}, \mathsf{df} = \mathsf{n} - 1)$

Tabella 13.5: Test unilaterale destro per una normale con significatività α

Test unilaterale destro	Regione di accettazione di $\mathbf{H_0}$
Test su μ con varianza σ^2 nota $\mathbf{H}_0 : \mu \geq \mu_0, \quad \mathbf{H}_1 : \mu < \mu_0$	$\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} > -z_\alpha$ $-z_\alpha$ si calcola con $\mathsf{qnorm}(\mathsf{alpha}, \mathsf{mean} = 0, \mathsf{sd} = 1)$ $pvalue = P(Z_n \leq z_{os})$ il p-value si calcola con $\mathsf{pnorm}(\mathsf{z_{os}}, \mathsf{mean} = 0, \mathsf{sd} = 1)$
Test su μ con varianza non nota $\mathbf{H}_0 : \mu \geq \mu_0, \quad \mathbf{H}_1 : \mu < \mu_0$	$\frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} > -t_{\alpha,n-1}$ $-t_{\alpha,n-1}$ si calcola con $\mathsf{qt}(\mathsf{alpha}, \mathsf{df} = \mathsf{n} - 1))$ $pvalue = P(T_n \leq t_{os})$ il p-value si calcola con $\mathsf{pt}(\mathsf{t_{os}}, \mathsf{df} = \mathsf{n} - 1)$
Test su σ^2 con valore medio μ noto $\mathbf{H}_0 : \sigma^2 \geq \sigma_0^2, \quad \mathbf{H}_1 : \sigma^2 < \sigma_0^2$	$\sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma_0} \right)^2 > \chi^2_{1-\alpha,n}$ $\chi^2_{1-\alpha,n}$ si calcola con $\mathsf{qchisq}(\mathsf{alpha}, \mathsf{df} = \mathsf{n})$
Test su σ^2 con valore medio non noto $\mathbf{H}_0 : \sigma^2 \geq \sigma_0^2, \quad \mathbf{H}_1 : \sigma^2 < \sigma_0^2$	$\frac{(n-1)\, s_n^2}{\sigma_0^2} > \chi^2_{1-\alpha,n-1}$ $\chi^2_{1-\alpha,n-1}$ si calcola con $\mathsf{qchisq}(\mathsf{alpha}, \mathsf{df} = \mathsf{n} - 1)$

Tabella 13.6: Test sulla media con significatività α per campioni numerosi

	Regione di accettazione di $\mathbf{H_0}$
Test bilaterale $\mathbf{H}_0 : \mu = \mu_0, \quad \mathbf{H}_1 : \mu \neq \mu_0$	$-z_{\alpha/2} < \frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} < z_{\alpha/2}$ $z_{\alpha/2}$ si calcola con $\mathsf{qnorm}(1 - \mathsf{alpha}/2, \mathsf{mean} = 0, \mathsf{sd} = 1)$
Test unilaterale sinistro $\mathbf{H}_0 : \mu \leq \mu_0, \quad \mathbf{H}_1 : \mu > \mu_0$	$\frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} < z_\alpha$ z_α si calcola con $\mathsf{qnorm}(1 - \mathsf{alpha}, \mathsf{mean} = 0, \mathsf{sd} = 1)$
Test unilaterale destro $\mathbf{H}_0 : \mu \geq \mu_0, \quad \mathbf{H}_1 : \mu < \mu_0$	$\frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} > -z_\alpha$ $-z_\alpha$ si calcola con $\mathsf{qnorm}(\mathsf{alpha}, \mathsf{mean} = 0, \mathsf{sd} = 1)$

Capitolo 14

Criterio del chi-quadrato

In questo capitolo dedicheremo l'attenzione al *criterio di verifica delle ipotesi del chi-quadrato*. Ci siamo finora occupati di ricavare informazioni da un campione estratto da una popolazione descritta da una variabile aleatoria X caratterizzata da una funzione di probabilità (nel caso discreto) o densità di probabilità (nel caso continuo) $f(x, \vartheta)$, stimando il parametro non noto ϑ (o i parametri non noti) della popolazione con stime puntuali ed intervallari. Abbiamo inoltre considerato il problema della verifica delle ipotesi statistiche considerando test unilaterali e bilaterali.

In molti problemi reali, si desidera verificare se il *campione osservato può essere stato estratto da una popolazione descritta da una variabile aleatoria X con funzione di distribuzione $F_X(x)$*. A questo scopo, utilizzeremo il *criterio di verifica delle ipotesi del chi–quadrato*, detto anche *test del chi–quadrato* o *test del buon adattamento*.

14.1 Criterio del chi–quadrato bilaterale

Con il criterio del chi–quadrato si desidera verificare l'ipotesi che un certa popolazione, descritta da una variabile aleatoria X, sia caratterizzata da una funzione di distribuzione $F_X(x)$, con k parametri non noti da stimare.

Denotando con $\mathbf{H}_0$ l'ipotesi soggetta a verifica (*ipotesi nulla*) e con $\mathbf{H}_1$ l'*ipotesi alternativa*, il test chi–quadrato con livello di significatività α mira a verificare l'ipotesi nulla

> $\mathbf{H_0}$*: X ha una funzione di distribuzione $F_X(x)$ (avendo stimato k parametri non noti in base al campione)*

in alternativa all'ipotesi

> $\mathbf{H_1}$*: X non ha una funzione di distribuzione $F_X(x)$,*

dove α è la *probabilità massima di rifiutare l'ipotesi nulla quando essa è vera*.

Occorre determinare un test ψ con livello di significatività α che permetta di determinare una regione di accettazione e di rifiuto dell'ipotesi nulla. Il test di verifica delle ipotesi considerato è bilaterale (o a due code).

Suddividiamo l'insieme dei valori che la variabile aleatoria X può assumere in r sottoinsiemi $I_1, I_2, \ldots, I_r$ (classi o categorie) in modo che risulti essere uguale a p_i la probabilità che, secondo la distribuzione ipotizzata, la variabile aleatoria assuma un valore appartenente a I_i, ossia

$$p_i = P(X \in I_i) \qquad (i = 1, 2, \ldots, r). \tag{14.1}$$

Si estrae poi un campione $x_1, x_2, \ldots, x_n$ di ampiezza n e si osservano le frequenze assolute $n_1, n_2, \ldots, n_r$ con cui gli n elementi si distribuiscono nei rispettivi insiemi $I_1, I_2, \ldots, I_r$. Quindi n_i rappresenta il *numero degli elementi del campione* che cadono nell'intervallo I_i $(i = 1, 2, \ldots, r)$. È chiaro che

$$\begin{aligned} &p_i \geq 0 \quad (i = 1, 2, \ldots, r), \qquad \sum_{i=1}^{r} p_i = 1; \\ &n_i \geq 0 \quad (i = 1, 2, \ldots, r), \qquad \sum_{i=1}^{r} n_i = n. \end{aligned} \tag{14.2}$$

Si nota che la probabilità che esattamente n_1 elementi appartengano ad I_1, n_2 elementi appartengano ad I_2, ..., n_r elementi appartengano ad I_r è

$$p(n_1, n_2, \ldots, n_r) = \frac{n!}{n_1!\, n_2! \ldots n_r!}\, p_1^{n_1} p_2^{n_2} \ldots p_r^{n_r}, \tag{14.3}$$

ossia una funzione di probabilità multinomiale. Ne segue che il numero medio di elementi che cadono nell'intervallo I_i è $n\, p_i$.
Si calcola poi la quantità

$$\chi^2 = \sum_{i=1}^{r} \left(\frac{n_i - n\, p_i}{\sqrt{n\, p_i}} \right)^2. \tag{14.4}$$

Il criterio chi–quadrato si basa sulla statistica

$$Q = \sum_{i=1}^{r} \left(\frac{N_i - n\, p_i}{\sqrt{n\, p_i}} \right)^2, \tag{14.5}$$

dove N_i è la variabile aleatoria che descrive il numero degli elementi del campione casuale $X_1, X_2, \ldots, X_n$ (costituito da n variabili aleatorie osservabili, indipendenti e identicamente distribuite con la stessa legge di probabilità $F_X(x)$ della popolazione) che cadono nell'intervallo I_i $(i = 1, 2, \ldots, r)$.
Se la variabile aleatoria X ha una funzione di distribuzione $F_X(x)$ con k parametri non noti, si può dimostrare che per n sufficientemente grande la funzione di distribuzione della statistica Q è approssimabile con la funzione di distribuzione chi–quadrato con $r - k - 1$ gradi di libertà. Si sottrae 1 da r a causa della prima delle condizioni (14.2) secondo la quale se conosciamo $r - 1$ delle probabilità p_i la rimanente probabilità può essere univocamente determinata e si sottrae k

poiché si suppone che siano k i parametri indipendenti non noti sostituiti da stime.

Per garantire che ogni classe contenga in media almeno 5 elementi, si ritiene valida l'approssimazione se risulta

$$\min(np_1, np_2, \ldots, np_r) \geq 5. \tag{14.6}$$

Nella Figura 14.1 è rappresentata la densità chi-quadrato con $r-k-1$ gradi di libertà e sono sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla del test bilaterale considerato.

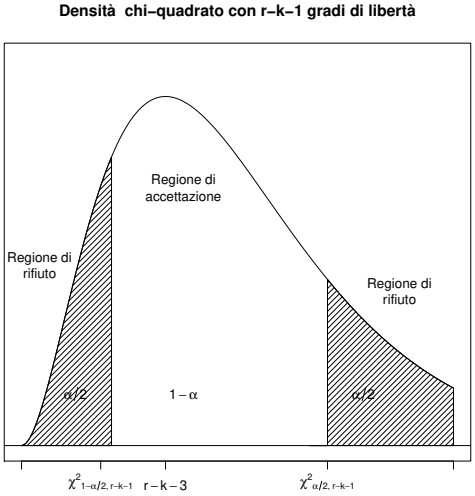

Figura 14.1: Zone di accettazione e di rifiuto dell'ipotesi nulla del test chi–quadrato bilaterale.

Si giunge così alla definizione del *test chi–quadrato bilaterale*.

Proposizione 14.1 *Per un campione sufficientemente numeroso di ampiezza n, il test chi–quadrato bilaterale di misura α è il seguente:*

- *si accetti l'ipotesi H_0 se $\chi^2_{1-\alpha/2,r-k-1} < \chi^2 < \chi^2_{\alpha/2,r-k-1}$,*
- *si rifiuti l'ipotesi H_0 se $\chi^2 < \chi^2_{1-\alpha/2,r-k-1}$ oppure $\chi^2 > \chi^2_{\alpha/2,r-k-1}$*

dove $\chi^2_{\alpha/2,r-k-1}$ e $\chi^2_{1-\alpha/2,r-k-1}$ sono soluzioni delle equazioni:

$$P(Q < \chi^2_{1-\alpha/2,r-k-1}) = \frac{\alpha}{2}, \qquad P(Q < \chi^2_{\alpha/2,r-k-1}) = 1 - \frac{\alpha}{2}. \tag{14.7}$$

Nel prossimo paragrafo applichiamo il criterio del chi-quadrato ipotizzando che il campione provenga da una popolazione di Poisson e da una popolazione normale.

14.2 Applicazioni

Esempio 14.1 (Poisson) In un incrocio stradale sono stati registrati il numero di incidenti che si sono verificati ogni giorno per un totale di 75 giorni distinti. I risultati sono

```
> camppois<-c(0, 3, 2, 0, 1, 2, 1, 1, 0, 1, 0, 1, 0, 0, 0,
+ 0, 0, 1, 0, 2, 0, 1, 0, 0, 0, 0, 1, 1, 3, 2,
+ 0, 1, 0, 1, 1, 0, 2, 3, 2, 1, 0, 0, 0, 1, 0,
+ 0, 0, 1, 0, 3, 0, 1, 0, 2, 4, 2, 0, 1, 1, 3,
+ 1, 0, 1, 0, 0, 0, 1, 0, 2, 4, 2, 0, 1, 2, 3)
>
> n<-length(camppois)
> n
[1] 75
>
> freq<-table(camppois)
> freq
camppois
 0  1  2  3  4
34 22 11  6  2
```

In questo caso, l'ampiezza del campione è $n = 75$ e corrisponde al numero di giorni considerati.

Si nota che nei 75 giorni nell'incrocio stradale in esame si sono verificati: 0 incidenti in 34 giorni, 1 incidente in 22 giorni, 2 incidenti in 11 giorni, 3 incidenti in 6 giorni e 4 incidenti in 2 giorni.

Si desidera verificare se il numero di incidenti sia descrivibile con una variabile aleatoria X di Poisson di parametro λ, ossia:

$$p_X(x) = \frac{\lambda^x}{x!}\, e^{-\lambda} \qquad (x = 0, 1, \ldots).$$

con $\lambda > 0$. I dati del campione permettono di ottenere una stima del parametro λ. Infatti, ricordando che uno stimatore corretto con varianza uniformemente minima del parametro λ di una distribuzione di Poisson risulta essere la media campionaria, si ha:

```
> stimalambda<-mean(camppois)
> stimalambda
[1] 0.9333333
```

Supponiamo di considerare 4 categorie corrispondenti agli intervalli $I_1 = \{0\}$, $I_2 = (0, 1]$, $I_3 = (1, 2]$, $I_4 = (2, +\infty)$. Le probabilità associate agli intervalli $p_1 = p_X(0)$, $p_2 = p_X(1)$, $p_3 = p_X(2)$ e $p_4 = 1 - p_X(0) - p_X(1) - p_X(2)$ possono essere così calcolate:

```
> p<-numeric(4)
> p[1]<-dpois(0,stimalambda)
> p[2]<-dpois(1,stimalambda)
> p[3]<-dpois(2,stimalambda)
> p[4]<-1-p[1]-p[2]-p[3]
> p
[1] 0.39324072 0.36702467 0.17127818 0.06845643
```

```
>
> sum(p)
[1] 1
```

Si nota che $p_1 + p_2 + p_3 + p_4 = 1$. Essendo

```
> min(n*p[1],n*p[2],n*p[3],n*p[4])
[1] 5.134232
```

maggiore di 5, la condizione (14.6) è soddisfatta. Il numero di elementi del campione appartenente ai quattro intervalli è

```
> r<-4
> nint<-numeric(r)
> nint[1]<-length(which(camppois==0))
> nint[2]<-length(which(camppois==1))
> nint[3]<-length(which(camppois==2))
> nint[4]<-length(which(camppois>2))
> nint
[1] 34 22 11  8
> sum(nint)
[1] 75
```

Calcoliamo ora χ^2 definito in (14.4)

```
> chi2<-sum(((nint-n*p)/sqrt(n*p))^2)
> chi2
[1] 3.663227
```

ossia $\chi^2 = 3.66$. In questo caso il numero di categorie è $r = 4$ e occorre porre $k = 1$ poiché la probabilità di Poisson contiene un parametro non noto. Pertanto, si ha $r - k - 1 = 2$ e scegliendo $\alpha = 0.01$ occorre calcolare $\chi^2_{1-\alpha/2,2}$ e $\chi^2_{\alpha/2,2}$:

```
> r<-4
> k<-1
> alpha<-0.01
> qchisq(alpha/2,df=r-k-1)
[1] 0.01002508
> qchisq(1-alpha/2,df=r-k-1)
[1] 10.59663
```

da cui segue che $\chi^2_{1-\alpha/2,r-k-1} = 0.010$ e $\chi^2_{\alpha/2,r-k-1} = 10.597$. Essendo $0.010 < \chi^2 < 10.597$, l'ipotesi H_0 di popolazione di Poisson può essere accettata. $\Diamond$

Esempio 14.2 (Normale) 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)
```

```
>
> n<-length(campnorm)
> n
[1] 50
>
> m<-mean(campnorm)
> m
[1] 121.872
> d<-sd(campnorm)
> d
[1] 6.735469
```

Si nota che la media campionaria $\overline{x} = 121.872\, m^2$ e la deviazione standard campionaria è $s = 6.735\, m^2$.

Applicando il test chi–quadrato di misura $\alpha = 0.05$, *si desidera verificare se la popolazione da cui proviene il campione può essere descritta da una variabile aleatoria X di densità normale*

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

Supponiamo di suddividere l'insieme dei valori che tale variabile aleatoria normale X può assumere in $r = 5$ sottoinsiemi $I_1, I_2, \ldots, I_5$ in modo che risulti essere uguale a $p_i = 0.2$ la probabilità che X assuma un valore appartenente a I_i $(i = 1, 2, \ldots, 5)$. La condizione (14.6) è verificata essendo $n\, p_i = 50 \cdot 0.2 = 10 \geq 5$. Ricordando che uno stimatore di μ è la media campionaria e uno stimatore di σ^2 è la varianza campionaria, utilizzando i quantili della distribuzione normale possiamo determinare i sottoinsiemi $I_1, I_2, \ldots, I_5$

```
> a<-numeric(4)
> for(i in 1:4)
+ a[i]<-qnorm(0.2*i,mean=m,sd=d)
> a
[1] 116.2033 120.1656 123.5784 127.5407
```

Gli intervalli $I_1, I_2, \ldots, I_5$ sono:

$$I_1 = (-\infty, 116.20), \quad I_2 = [116.2, 120.17), \quad I_3 = [120.17, 123.58),$$
$$I_4 = [123.58, 127.54), \quad I_5 = [127.54, +\infty).$$

Occorre ora determinare il numero di elementi del campione che cadono negli intervalli $I_1, I_2, \ldots, I_5$:

```
> r<-5
> nint<-numeric(r)
> nint[1]<-length(which(campnorm<a[1]))
> nint[2]<-length(which((campnorm>=a[1])&(campnorm<a[2])))
> nint[3]<-length(which((campnorm>=a[2])&(campnorm<a[3])))
> nint[4]<-length(which((campnorm>=a[3])&(campnorm<a[4])))
> nint[5]<-length(which(campnorm>=a[4]))
> nint
[1] 10 11  5 16  8
> sum(nint)
[1] 50
```

Segue che $n_1 = 10$, $n_2 = 11$, $n_3 = 5$, $n_4 = 16$ e $n_5 = 8$. Calcoliamo ora χ^2 definito in (14.4)

```
> chi2<-sum(((nint-n*0.2)/sqrt(n*0.2))^2)
> chi2
[1] 6.6
```

ossia $\chi^2 = 6.6$.

La distribuzione normale ha due parametri non noti (μ, σ^2) e quindi $k = 2$. Pertanto, la funzione di distribuzione della statistica Q è approssimabile con la funzione di distribuzione chi–quadrato con $r - k - 1 = 2$ gradi di libertà. Occorre quindi calcolare $\chi^2_{\alpha/2,2}$ e $\chi^2_{1-\alpha/2,2}$ con $\alpha = 0.05$.

```
> r<-5
> k<-2
> alpha<-0.05
> qchisq(alpha/2,df=r-k-1)
[1] 0.05063562
> qchisq(1-alpha/2,df=r-k-1)
[1] 7.377759
```

da cui segue che $\chi^2_{1-\alpha/2,r-k-1} = 0.0506$ e $\chi^2_{\alpha/2,r-k-1} = 7.378$. Essendo $0.0506 < \chi^2 < 7.378$, l'ipotesi H_0 di popolazione normale può essere accettata. $\Diamond$

Vi auguro di completare con serenità la vostra carriera universitaria e di inserirvi con successo nel mondo del lavoro.

Amelia G. Nobile

Indice