6 Statistica inferenziale

In questo capitolo si introduce la statistica inferenziale. Il *collegamento* fra la statistica descrittiva e la probabilità che abbiamo visto nei capitoli precedenti e la statistica inferenziale è rappresentato dalla *distribuzine dei campioni*.

Infatti, se l'indagine statistica deve essere fatta su una popolazione troppo grande per essere esaminata elemento per elemento, di questa popolazione se ne considerano dei *campioni* casuali e analizzando questi campioni si deducono informazioni sulla popolazione a cui appartengono. Affinchè il metodo abbia successo, il campione deve essere *rappresentativo*. Un modo efficiente per sceglierlo è quello di prenderlo in modo *casuale* dalla popolazione.

L'obiettivo della *statistica inferenziale*, come dice il nome stesso, è quello di conoscere la distribuzione di probabilità associata alla popolazione (*distribuzione della popolazione*) a partire dalla distribuzione di probabilità del campione (*distribuzione del campione*). Oggetto del presente capitolo è analizzare le relazioni esistenti fra le due distribuzioni, che permettono appunto di conoscere le caratteristiche della popolazione a partire da quelle del campione. Ci sono due modi per avere informazioni sulla statistica dei campioni. Uno richiede calcoli basati su regole di probabilità, l'altro è basato su simulazioni.

6.1 Distribuzione di campioni

Se $X_1, X_2, \ldots X_n$ sono variabili aleatorie indipendenti con $X_i \sim f$ allora si dice che $X_1, X_2, \ldots X_n$ sono un campione casuale semplice di dimensione n della popolazione f e si denota con SRS(n) (Simple Random Sample di dimensione n).

Definiamo ora che cos'è una statistica per poi studiare le statistiche dei SRS(n).

Una *statistica* è ogni quantità il cui valore può essere calcolato da campioni di dati.

La statistica è quindi una variabile aleatoria e viene solitamente indicata con lettere maiuscole. Il valore calcolato o un valore osservato della statistica viene invece denotato con lettera minuscola.

Per esempio la media aritmetica vista come statistica (prima di scegliere un particolare campione su cui calcolarla) si può denotare con $\bar{X}$ e il suo valore calcolato con $\bar{x}$. Ogni statistica, essendo una variabile casuale, ha una distribuzione di probabilità che è detta *distribuzione di campioni* per dire come la statistica varia in valore al variare del campione selezionato.

Se $X_1, X_2, \ldots X_n$ sono un SRS(n) di una popolazione con distribuzione f con media μ e deviazione standard σ, allora la media e la deviazione standard di $\bar{X}$ (media di X) sono date da: $\mu_{\bar{X}} = \mu, \ \sigma_{\bar{X}} = \sigma/\sqrt{n}$.

6.1.1 Statistiche di campioni da una distribuzione normale

6.1.1.1 Distribuzione della media campionaria

Supponiamo ora di estrarre un campione SRS(n) da una distribuzione normale. Vediamo come è distribuita la media di questo campione al variare del campione stesso. Vale il seguente risultato.

Siano $X_1, X_2, \ldots X_n$ un SRS(n) da una distribuzione normale con media μ e deviazione standard σ. Allora la variabile aleatoria data dalla media campionaria $\bar{X}$ ha una distribuzione: `norm(mean=mu, sd=sigma/`$\sqrt{n}$`)`.

Facciamo una simulazione in R.

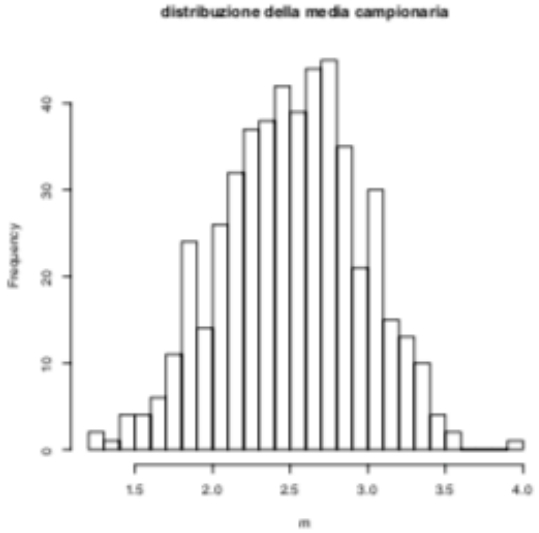
Esempio 6.1 *Consideriamo k campioni SRS(n), ciascuno estratto da una distribuzione normale con media e deviazione standard assegnate. Verifichiamo che la distribuzione della media dei campioni è ancora una distribuzione normale con le caratteristiche riportate precedentemente.*

Soluzione.

```
> mu <- 2.5
> sigma <- 1
> k <- 500
> n <- 5
> m <- replicate(k, mean(rnorm(n, mu, sigma)))
> mean(m)

## [1] 2.512343

> hist(m, breaks = 20,
+      main = "distribuzione della media campionaria")
```



In questo caso la media della popolazione è 2.5 e la media campionaria è 2.510042, quindi una approssimazione esatta alla prima cifra decimale. Se ripetiamo la simulazione aumentando il valore di n e di k, il valore della media campionaria sarà sempre più vicino a quello della media esatta. Inoltre, si può verificare come la distribuzione della media campionaria si avvicini sempre più alla normale con media μ e deviazione standard $\sigma/\sqrt{n}$ sempre all'aumentare di n e di k.

Dal risultato precedente segue immediatamente che se $X_1, X_2, \ldots X_n$ sono un SRS(n) da una distribuzione `(mean=mu, sd=sigma)`, allora la variabile aleatoria

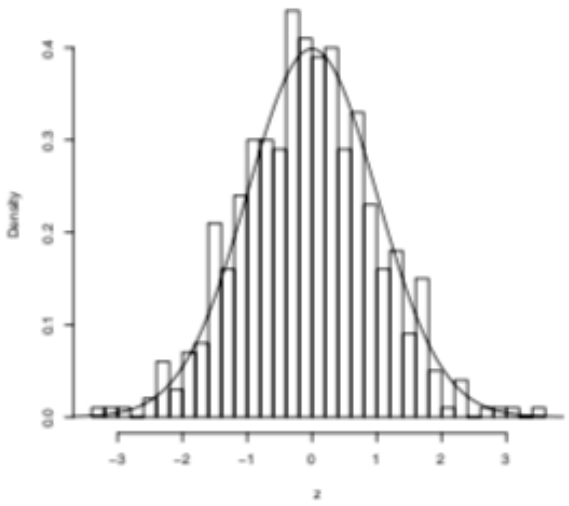
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

ha distribuzione campionaria normale standard.

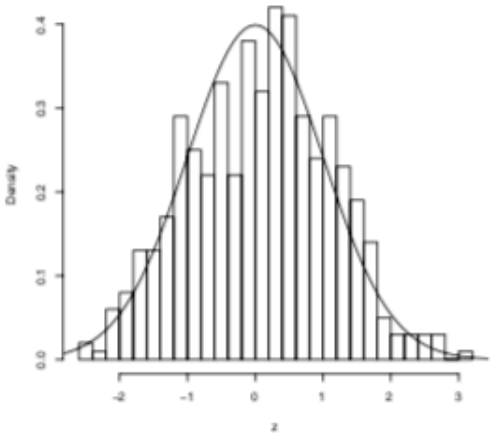
Esempio 6.2 *Ripetiamo ora una simulazione analoga alla precedente, in cui però visualizziamo l'istogramma della variabile aletaoria Z anzichè di $\bar{X}$ (nell'istogramma, viene disegnato anche il grafico della distribuzione normale standard per confronto). Verifichiamo che Z ha una distribuzione normale standard (cioè con media zero e deviazione standard 1). La simulazione viene ripetuta 2 volte, con $n = 5$ e $n = 15$ elementi per campione e con $k = 500$ e $k = 2000$ campioni.*

Soluzione.

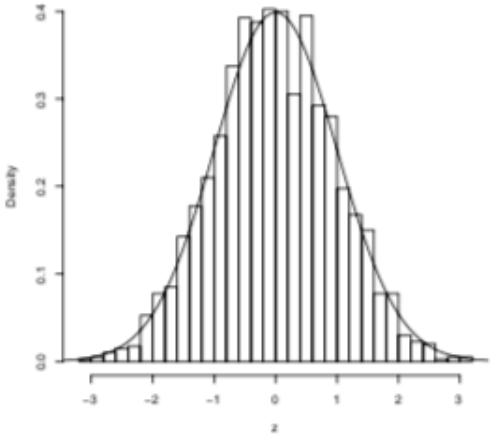
```
> mu <- 2.5
> sigma <- 1
> k <- 500
> n <- 5
> m <- replicate(k,mean(rnorm(n,mu,sigma)))
> s <- sigma/sqrt(n)
> z <- (m-mu)/s
> hist(z,breaks=30,main="",freq=FALSE)
> curve(dnorm(x),from=-4,to=4,add=TRUE)
```




```
>  n <- 15
>  m <- replicate(k,mean(rnorm(n,mu,sigma)))
>  s <- sigma/sqrt(n)
>  z <- (m-mu)/s
>  hist(z,breaks=30,main="",freq=FALSE)
>  curve(dnorm(x),from=-4,to=4,add <- TRUE)
```




```
> k <- 2000
>  m <- replicate(k,mean(rnorm(n,mu,sigma)))
>  s <- sigma/sqrt(n)
>  z <- (m-mu)/s
>  hist(z,breaks=30,main="",freq=FALSE)
>  curve(dnorm(x),from=-4,to=4,add=TRUE)
```



6.1.1.2 Distribuzione della varianza campionaria

Consideriamo ora la distribuzione della variabile aleatoria data dalla varianza campionaria. Nel caso della varianza vale il seguente risultato.

Siano $X_1, X_2, \ldots X_n$ un SRS(n) da una distribuzione normale con media μ e deviazione standard σ (`norm(mean = mu, sd = sigma)`) e sia S^2 la varianza campionaria:

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

Allora la varianza campionaria scalata:

$$\frac{n-1}{\sigma^2}S^2$$

ha distribuzione campionaria χ^2 con $n-1$ gradi di libertà (`chisq(df = n - 1)`).

Ne consegue che se $X_1, X_2, \ldots X_n$ sono un SRS(n) da una distribuzione `norm(mean = mu, sd = sigma)`, allora la variabile aleatoria

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

ha distribuzione t di student con $n-1$ gradi di libertà (`t(df = n - 1)`).

Quindi quando consideriamo campioni da una distribuzione normale possiamo avere con precisione informazioni sulle statistiche campionarie di media e varianza.

6.1.2 Il Teorema del limite centrale

Consideriamo ora il caso di campioni estratti da una popolazione con una distribuzione **diversa da quella normale**, di cui però sono note media e deviazione standard. Cosa possiamo dire delle statistiche campionarie (media, varianza, ...)?

La risposta a questa domanda è data dal *Teorema del limite centrale*.

Teorema del limite centrale. Siano $X_1, X_2, \ldots X_n$ una SRS(n) da una distribuzione di popolazione con media μ e deviazione standard σ. Allora la variabile aleatoria

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

ha una distribuzione campionaria che **ha come limite**, per $n \to \infty$, la distribuzione normale standard (`norm(mean = 0, sd = 1)`).

Questo teorema permette di avere quindi informazioni *per ogni distribuzione che abbia una deviazione standard finita*, almeno quando il numero di elementi nel campione è *sufficientemente grande*.

Quantificare la frase *sufficientemente grande* è difficile. Dipende, per esempio, dalla forma della distribuzione della popolazione da cui viene estratto il campione. Se

la distribuzione è simmetrica e mesocurtica sono sufficienti in generale pochi elementi (n può avere valori bassi come 5-6), mentre per distribuzioni fortemente asimmetriche e con curtosi pronunciate n deve avere valori più grandi affinchè la distribuzione campionaria approssimi la normale.

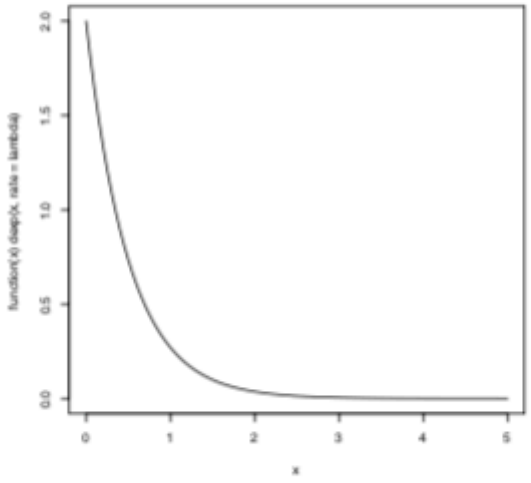
Vediamo un esempio di simulazione in R .

Esempio 6.3 *Consideriamo un campione SRS(n) da una distribuzione molto asimmetrica come quella esponenziale e verifichiamo su di esso il teorema del limite centrale. Consideriamo 3 campioni di 5, 15, 25 elementi rispettivamente e per ognuno di esso visualizzaimo l'istogramma della variabile aleatoria Z. Sullo stesso grafico è disegnata, per confronto, la curva della distribuzione normale a cui tende Z. In ognuno dei 3 casi sono presi $k = 500$ campioni.*

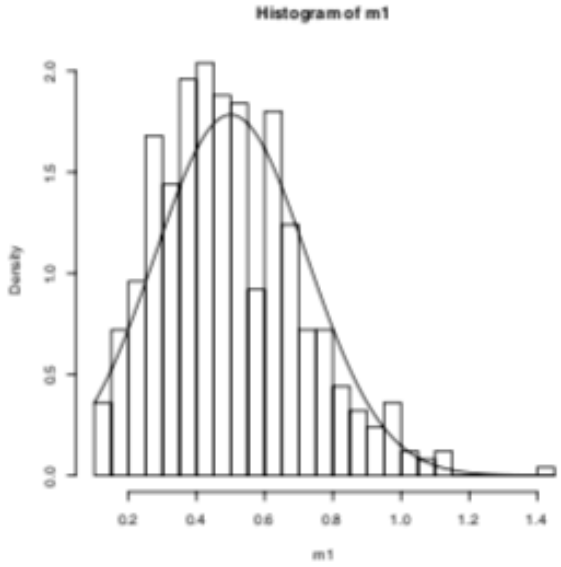
```
> k=500
> n1=5
> n2=15
> n3=35
> 'distribuzione di partenza'

## [1] "distribuzione di partenza"

> lambda=2
> plot(function(x) dexp(x,rate=lambda), 0, 5)
```



```
> m1=replicate(k,mean(rexp(n1,rate=lambda)))
> mu_ex=1/lambda
> sigma_ex=1/lambda
> hist(m1,breaks=30,freq=FALSE)
> s=sigma_ex/sqrt(n1)
> curve(dnorm(x,mu_ex,s),add=TRUE)
```

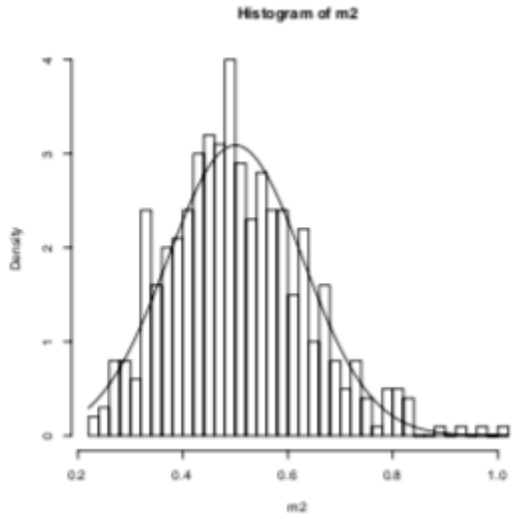


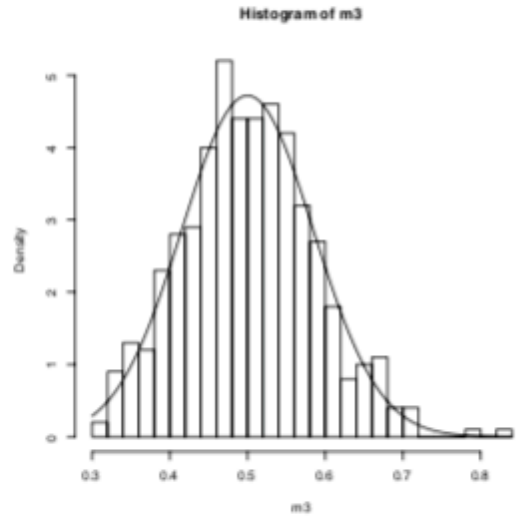

```
> m2=replicate(k,mean(rexp(n2,rate=lambda)))
> hist(m2,breaks=30,freq=FALSE)
> s=sigma_ex/sqrt(n2)
> curve(dnorm(x,mu_ex,s),add=TRUE)
```

```
> m3=replicate(k,mean(rexp(n3,rate=lambda)))
> hist(m3,breaks=30,freq=FALSE)
> s=sigma_ex/sqrt(n3)
> curve(dnorm(x,mu_ex,s),add=TRUE)
```

```
> show(mean(m3))

## [1] 0.5008415
```


Histogram of m2
Density
0
1
2
3
4
0.2
0.4
0.6
0.8
1.0
m2


Histogram of m3
Density
0
1
2
3
4
5
0.3
0.4
0.5
0.6
0.7
0.8
m3

6.2 Stime

6.2.1 Concetti generali

In questo paragrafo introduciamo le stime puntuali e le stime di intervalli. Innanzitutto precisiamo che la stima puntuale è relativa ad un parametro di interesse della popolazione, come può essere la media o altri parametri che caratterizzano la distribuzione.

La *stima puntuale di un parametro* θ (con le lettere greche indichiamo il parametro di interesse) è un numero che può essere un valore sensibile di θ.

Una stima puntuale è ottenuta scegliendo un'opportuna statistica e calcolando il suo valore a partire da campioni casuali. La statistica scelta è detta *stimatore puntuale di* θ e si indica generalmente con $\hat{\theta}$.

La stima di un parametro però non fornisce da sola sufficienti informazioni riguardo alla sua affidabilità. È quindi necssario affiancare alla stima puntuale la stima di un intervallo di valori possibili, detto *intervallo di confidenza*, ottenuto a partire da valori che 'misurano' il *grado di affidabilità* della stima. Più piccolo è l'intervallo stesso e più affidabile è la stima. Un grande intervallo di confidenza è segno di incertezza nella stima calcolata.

6.2.2 Stime puntuali: stimatori non distorti

Supponiamo di voler stimare il parametro θ con lo stimatore $\hat{\theta}$. Il meglio che si possa ottenere sarebbe che $\hat{\theta} = \theta$ per ogni campione considerato. In realtà, $\hat{\theta}$ è una variabile aleatoria, quindi può capitare che per un certo campione $\hat{\theta} < \theta$ e per un altro $\hat{\theta} > \theta$. In generale, si ha che:

$$\hat{\theta} = \theta + \text{errore di stima.}$$

Ovviamente, più è piccolo l'errore, migliore è l'estimatore.

Per valutare l'errore commesso nella stima, si calcolano delle misure di errore come, per esempio, l'errore quadratico medio (Mean Square Error) fra il valore stimato $\hat{\theta}$ e il valore esatto θ:

$$MSE = \frac{\sum_{i=1}^{n}(\hat{\theta} - \theta)^2}{n}.$$

Nel confronto fra due stimatori $\hat{\theta}_1$ e $\hat{\theta}_2$, viene considerato migliore quello che ha l'errore quadratico medio minore. È difficile trovare uno stimatore che sia sempre migliore degli altri per ogni θ, visto che lo stimatore stesso dipende da θ.

Uno stimatore puntuale $\hat{\theta}$ è detto *non distorto* se, detta $E(\hat{\theta})$ la media della variabile aleatoria $\hat{\theta}$, $E(\hat{\theta}) = \theta$ per ogni possibile valore di θ.

Se $\hat{\theta}$ è distorto, allora la differenza $E(\hat{\theta}) - E(\theta)$ si dice la *distorsione* di $\hat{\theta}$.

Da questa definizione si deduce che $\hat{\theta}$ è non distorto se la sua distribuzione di probabilità è sempre 'centrata' nel valore *vero* del parametro θ. Per esempio, se $\hat{\theta}$ è non distorto e $\theta = 100$, allora la distribuzione di probabilità di $\hat{\theta}$ deve avere media $\mu = E(\hat{\theta}) = 100$. Può sembrare necessario conoscere il valore di θ per sapere se uno stimatore è distorto oppure no; in realtà la distorsività è una proprietà dello stimatore indipendente dai valori θ.

Qual è l'importanza della proprietà di distorsione dello stimatore?

Principio della stima non distorta

Quando si deve scegliere fra diversi stimatori, scegliere quello *non distorto*.

Se ci sono più stimatori non distorti di uno stesso parametro allora si sceglie in base alla varianza dello stimatore.

In particolare:

Principio della varianza minima. Fra tutti gli stimatori non distorti di θ, scegliere quello con varianza minima. Il risultante stimatore $\hat{\theta}$ è detto *stimatore non distorto di minima varianza* (MVUE).

6.2.3 Stimatori non distorti di media e varianza

Siano $X_1, X_2, \ldots X_n$ campioni casuali da una distribuzione con media μ e varianza σ^2. Allora lo stimatore:

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

è non distorto per stimare la varianza σ^2. Lo stimatore che ha come denominatore n:

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

è distorto e la sua distorsione è $(n-1)/n\sigma^2 - \sigma^2 = -\sigma^2/n$. Essendo la distorsione di P negativa, lo stimatore P tende a sottostimare la varianza σ^2.

Siano $X_1, X_2, \ldots X_n$ campioni casuali da una distribuzione con media μ. Allora lo stimatore $\bar{X}$ è *uno stimatore non distorto della media* μ.

In generale ci sono diversi stimatori della media. Per esempio si possono usare medie trimmate, oppure la media fra due osservazioni estreme. La bontà dello stimatore in questo caso dipende dalla distribuzione dalla quale sono stati estratti i campioni. Per esempio, se i campioni sono estratti da una distribuzione normale, vale il seguente risultato.

Siano $X_1, X_2, \ldots X_n$ SRS(n) da una distribuzione normale con media μ e deviazione standard σ. Allora l'estimatore $\bar{X}$ è l'estimatore MVUE della media.

6.2.4 Metodi per le stime puntuali

Come ottenere gli stimatori, quando si devono stimare parametri diversi da media e varianza? Ci sono principalmente due metodi:

1. Il *metodo dei momenti*. È meno costoso, ma i risultati sono meno accurati.
2. Il *metodo di Massima Verosimiglianza o Maximimum LIkelihood (ML)*. È più costoso ma i risultati sono migliori in generale.

6.2.4.1 Metodo dei momenti

Sia $X_1, X_2, \ldots X_n$ un campione casuale da una PDF o PMF $f_X(x)$. Per $k = 1, 2, 3, \ldots$ si definisce *k-esimo momento della distribuzione* $f_X(x)$ il valore $E(X^k)$.

Si definisce *k-esimo momento del campione* il valore:

$$1/n \sum_{i=1}^{n} X_i^k$$

.

Con il metodo dei momenti gli stimatori sono definiti nel modo seguente.

Sia $X_1, X_2, \ldots X_n$ un campione casuale da una distribuzione con PDF definita come $f_X(x; \theta_1, \theta_2, \ldots \theta_m)$ dove $\theta_1, \theta_2, \ldots \theta_m$ sono i parametri da stimare. Allora i *momenti estimatori* $\hat{\theta}_1, \hat{\theta}_2, \ldots \hat{\theta}_m$ sono ottenuti uguagliando i primi m momenti campioni ai corrispondenti momenti della distribuzione della popolazione e risolvendo rispetto a $\theta_1, \theta_2, \ldots \theta_m$.

Esempio 6.4 *Sia $X_1, X_2, \ldots X_n$ un SRS(n) di tempi di servizio a n clienti dove la distribuzione di probabilità è assunta esponenziale con parametro λ. Stimare con il metodo dei momenti λ.*

Soluzione.

Poiché in questo caso c'è un solo parametro da stimare, lo stimatore si calcola con il metodo dei momenti dall'equazione:

$$E(X) = \bar{X}$$

cioè:

$$E(X) = \frac{1}{\lambda}, \ \rightarrow \lambda = \frac{1}{\bar{X}}.$$

Quindi:

$$\hat{\lambda} = \frac{1}{\bar{X}}.$$

Esempio 6.5 *Sia $X_1, X_2, \ldots X_n$ un SRS(n) da una distribuzione* `gamma` (α, β). *Stimare con il metodo dei momenti α e β.*

Soluzione. In questo caso ci sono due parametri da stimare, quindi si devono uguagliare i primi due momenti.

$E(X) = \alpha\beta$, $E(X^2) = \beta^2(\alpha + 1)\alpha$.

Quindi impongo:

$$\bar{X} = \alpha\beta, \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \alpha(\alpha + 1)\beta^2.$$

Da cui ricavo:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \bar{X}^2 + \alpha\beta^2.$$

Dividendo entrami i membri della seconda equazione per i corrispondenti della prima equazione, si ottengono i seguenti stimatori:

$$\hat{\alpha} = \frac{\bar{X}^2}{1/n\sum X_i^2 - \bar{X}^2}, \quad \hat{\beta} = \frac{(1/n)\sum X_i^2 - \bar{X}^2}{\bar{X}}.$$

6.2.4.2 Metodo di Massima Verosimiglianza

Il metodo di Massima Verosimiglianza fu introdotto intorno al 1920. È utilizzato dagli statistici soprattutto per problemi di grandi dimensioni.

Il metodo è basato sulla massimizzazione di una funzione, detta *funzione di verosimiglianza*.

Supponiamo che $X_1, X_2, \ldots X_n$ sia un SRS(n) da una distribuzione con PDF o PMF $f(x_1, x_2, \ldots x_n; \theta_1, \theta_2, \ldots \theta_m)$. Se consideriamo f come funzione dipendente da m parametri $\theta_1, \theta_2, \ldots \theta_m$, assegnati i valori osservati $x_1, x_2, \ldots x_n$, la *funzione di verosimiglianza* è:

$$L(\theta_1, \theta_2, \ldots \theta_m) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, \ldots \theta_m)$$

Lo stimatore di Massima Verosimiglianza si ottiene sostituendo le variabili aleatorie $X_1, X_2, \ldots X_n$ alle osservazioni $x_1, x_2, \ldots x_n$ e calcolando il massimo della funzione L:

$$\hat{\theta}_1, \hat{\theta}_2, \ldots \hat{\theta}_m = argmax_{\theta_1, \theta_2, \ldots \theta_m} L(\theta_1, \theta_2, \ldots \theta_m).$$

Alcune considerazioni riguardo al calcolo degli stimatori MLE:

- Non è detto che esista il massimo della funzione di verosimiglianza e anche se esiste non è detto che sia unico.

- Spesso è più semplice minimizzare l'opposto del logaritmo naturale di $L(\theta)$, cioè $-ln(L(\theta)$, anzichè massimizzare $L(\theta)$. Poiché la funzione *logaritmo* è monotona, i due problemi hanno le stesse soluzioni (ricordiamo inoltre che $argmin_x f(x) = argmin_x - f(x)$ per qualsiasi funzione f).
- Spesso non è possibile calcolare il minimo della funzione $-ln(L(\theta))$ in modo esatto, quindi si ricorre ad algoritmi numerici di minimizzazione. Gli algoritmi numerici per la minimizzazione (o massimizzazione) di funzioni in una o più variabili sono tutti algoritmi iterativi. Gli algoritmi iterativi calcolano la soluzione esatta del problema al limite del procedimento (per $n \to \infty$, dove n è l'indice dell'iterazione); essendo impossibilie implementare un procedimento infinito, l'algoritmo viene arrestato ad una iterazione $k < \infty$ determinata dal criterio di arresto. L'algoritmo calcola quindi una soluzione approssimata con la precisione legata alla tolleranza assegnata (vedi **Capitolo 7**).

Esempio 6.6 *Siano $X_1, X_2, \ldots X_n$ un SRS(n) da una distribuzione esponenziale con parametro λ. Stimare λ con il metodo MLE.*

Soluzione.

Per l'indipendenza dei campioni, la funzione di verosimiglianza è il prodotto delle PDF:

$$L(\lambda) = \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2} \ldots \lambda e^{-\lambda x_n} = \lambda^n e^{-\lambda \sum x_i}$$

Il logaritmo naturale di f è:

$$ln(f(x_1, x-2, \ldots x_n; \lambda)) = n \cdot ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

Facendo

$$(\partial/\partial\lambda)(ln(f(x_1, x-2, \ldots x_n; \lambda))) = n/\lambda - \sum_{i=1}^{n} x_i = 0$$

perciò

$$\lambda = n/(\sum_{i=1}^{n})x_i = 1/\bar{x}.$$

Quindi lo stimatore di MLE è $1/\bar{X}$, che è lo stesso risultato ottenuto con il metodo dei momenti.

In R il calcolo di una stima MLE si può fare in diversi modi. Si può usare la funzione `optimize` del pacchetto `stats`, R Core Team [2015], che calcola il massimo o il minimo di una funzione in una variabile ($f : [a, b] \subset R \to R$).

Esempio 6.7 *Il programma R per calcolare numericamente la stima precedente è il seguente, sia calcolando il massimo della funzione di verosimiglianza, che il minimo del suo logaritmo.*

Soluzione.

```
> n <- 30
> x <- rexp(n, rate = 1.5)
> fl <- function(theta, x) prod(dexp(x, rate = theta))
> optimize(fl, interval = c(0, 2), x = x, maximum = TRUE)

## $maximum
## [1] 1.266997
##
## $objective
## [1] 1.13368e-10

> fll <- function(theta, x) -log(prod(dexp(x, rate = theta)))
> optimize(fll, interval = c(0, 2), x = x, maximum = FALSE)

## $minimum
## [1] 1.267006
##
## $objective
## [1] 22.90038
```

I parametri della funzione `optimize` sono, nell'ordine, i seguenti:

- `f` è la funzione da massimizzare (minimizzare) rispetto al primo argomento della funzione stessa
- `interval` è l'intervallo in cui cercare il massimo (minimo)
- Se al parametro `maximum` non è assegnato il valore `TRUE`, viene calcolato il minimo della funzione.

L'algoritmo utilizzato è un algoritmo iterativo (consultare l'`help` della funzione per maggiori dettagli). Per una più approfondita discussione riguardo agli algoritmi di ricerca di massimo o minimo di funzione si veda il **Capitolo 7**.

Il primo output è il punto di massimo della funzione di verosimiglianza, cioè la stima del parametro p, mentre il secondo output è il valore della funzione di massima verosimiglianza nel punto di massimo.

Esempio 6.8 *Supponiamo ora che $X_1, X_2, \ldots X_n$ siano un SRS(n) da una distribuzione normale con media μ e deviazione standard σ. Vogliamo stimare i parametri μ e σ con uno stimatore MLE.*

Soluzione.

La funzione di verosimiglianza è:

$$\begin{aligned} L(\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/\sigma^2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_2-\mu)^2/\sigma^2} \cdots \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/\sigma^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\sum_i (x_i-\mu)^2/\sigma^2}. \end{aligned}$$

Quindi:

$$ln(L(\mu, \sigma^2)) = -\frac{n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2.$$

Per calcolare il minimo, calcolo le derivate parziali e le annullo:

$$\frac{\partial}{\partial \mu} ln(L(\mu, \sigma^2)) = 0, \quad \frac{\partial}{\partial \sigma^2} ln(L(\mu, \sigma^2)) = 0,$$

risolvendo il sistema si ottengono i seguenti valori, che rappresentano le stime MLE per μ e σ^2, rispettivamente:

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \frac{\sum (X_i - \bar{X})^2}{n}.$$

Confrontando con le stime non distorte per la media e varianza della distribuzione normale, si ha che lo stimatore $\hat{\mu}$ risulta non distorto, mentre lo stimatore $\hat{\sigma}^2$ è distorto, in quanto il denominatore è uguale a n anzichè $n - 1$.

Esempio 6.9 *Si deve stimare il parametro di probabilità p di una distribuzione binomiale.*

Soluzione.

La funzione di verosimiglianza in questo caso è:

$$L(p) = p^{\sum_i x_i} (1-p)^{n - \sum_i x_i}$$

ottenuta come prodotto delle funzioni binomiali `binom(size = 1, prob = p)`.

Consideriamo un campione memorizzato nel file `cars` di R.

```
> x <- mtcars$am
> L <- function(p, x) prod(dbinom(x, size = 1, prob = p))
> optimize(L, interval = c(0, 1), x = x, maximum = TRUE)

## $maximum
## [1] 0.4062458
##
## $objective
## [1] 4.099989e-10
```

Come si era detto prima, numericamente è più conveniente calcolare il minimo dell'opposto del logaritmo:

```
> x <- mtcars$am
> menologL <- function(p, x)
+ {
+   -sum(dbinom(x, size = 1,
+         prob = p, log = TRUE))
+ }
> optimize(menologL, interval = c(0, 1), x = x)

## $minimum
## [1] 0.4062525
##
## $objective
## [1] 21.61487
```

Il punto di minimo è molto vicino al punto di massimo precedente (la differenza dipende da approssimazioni di calcolo con i numeri finiti) mentre il valore della funzione è ovviamente diverso.

Esempio 6.10 *Consideriamo il file* `PlantGrowth`*. È nel pacchetto* `datasets` *di R. È un data frame con due variabili (*`weigth` *e* `group`*) e 30 casi, che rappresentano le misure (come peso) relative alla crescita di piante dopo due tipi di trattamento. Per quanto riguarda la variabile* `weigth`*, supponiamo che i suoi valori siano variabili aleatorie* $X_1, X_2, \ldots X_n$ *da una distribuzione* `norm(`μ`, `σ`)`*. Vogliamo determinare una stima MLE di* $\theta = (\mu, \sigma^2)$*.*

Soluzione.

In questo caso i parametri da stimare sono due, quindi la funzione di verosimiglianza è una funzione $f : \Omega \subset R^2 \to R$.

In questo caso si può utilizzare la funzione R `mle` della libreria `stats4`, R Core Team [2015], che applica un metodo iterativo per calcolare il minimo di una funzione in più variabili. L'algoritmo richiede anche un iterato iniziale, dal quale dipende la convergenza del metodo stesso. Infatti, se ci sono minimi locali, l'iterato iniziale determina a quale di questi minimi converge l'algoritmo. Per questo, più l'iterato iniziale è vicino alla soluzione, maggiori sono le probabilità che la soluzione cercata sia proprio quella determinata dall'algoritmo.

```
> require(stats4)

## Loading required package:  stats4
```

```
> x <- PlantGrowth$weight
> menologL <- function(mu, sigma2)
+ {
+   -sum(dnorm(x, mean = mu,
+         sd = sqrt(sigma2), log = TRUE))
+ }
> mle_est <- mle(menologL,
+                start = list(mu = 5,
+                             sigma2 = 0.5))

> summary(mle_est)

## Maximum likelihood estimation
##
## Call:
## mle(minuslogl = menologL, start = list(mu = 5, sigma2 = 0.5))
##
## Coefficients:
##         Estimate Std. Error
## mu      5.0729848  0.1258666
## sigma2 0.4752721  0.1227108
##
## -2 log L: 62.82084
```

Il comando `summary` riporta alcune informazioni riguardo al risultato di una funzione di approssimazione di un modello, quale appunto la funzione `elm`. In particolare, riporta in questo caso i valori stimati dei parametri e *l'errore standard*, cioè la deviazione standard dello stimatore $\hat{\theta}$.

6.2.5 Intervalli di confidenza per la media: concetti e interpretazione

In questo paragrafo introduciamo il concetto di *intervallo di confidenza*, che ci permette di valutare probabilisticamente l'affidabilità della stima effettuata sul parametro.

6.2.5.1 Caso di popolazione con distribuzione normale

Consideriamo innanzitutto una situazione poco realistica, ma utile per introdurre il concetto di intervallo di confidenza che vogliamo studiare, cioè consideriamo campioni da una distribuzione normale di cui si conosce la deviazione standard ma non la media.

Dati $X_1, X_2, \ldots X_n$ SRS(n) da una distribuzione `norm(`μ`, `σ`)`, sia μ non noto. Possiamo stimare μ con $\bar{X}$ e abbiamo anche visto che questo coincide con la stima MLE.

Quanto è buona questa stima?

Sappiamo dal teorema del Limite Centrale che:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \texttt{norm(mean=0, sd=1)}$$

L'affidabilità della stima dipende dal livello di probabilità che vogliamo; si deve quindi scegliere un livello di probabilità, cioè un valore nell'intervallo $[0,1]$. Se consideriamo una probabilità $1-\alpha$ alta, per esempio 0.95 ($\alpha = 0.05$), possiamo calcolare il quantile $z_{\alpha/2}$. Allora si ha che:

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha.$$

Facendo i seguenti passaggi per trovare un intervallo di probabilità per la media μ:

$$-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2},$$

$$-z_{\alpha/2}\left(\sigma/\sqrt{n}\right) \leq \bar{X}-\mu \leq z_{\alpha/2}\left(\sigma/\sqrt{n}\right),$$

$$\bar{X}-z_{\alpha/2}\left(\sigma/\sqrt{n}\right) \leq \mu \leq \bar{X}+z_{\alpha/2}\left(\sigma/\sqrt{n}\right).$$

Perciò:

$$P\left(\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Definizione. L'intervallo:

$$\left(\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

è detto *intervallo di confidenza* $100(1-\alpha)\%$ *di* μ. Il valore $1-\alpha$ è detto *coefficiente o livello di confidenza*.

L'intervallo è casuale, poiché i due estremi dipendono dalla variabile aleatoria $\bar{X}$, ed è centrato su $\bar{X}$.

Possiamo anche dire che *la probabilità che il valore esatto* μ *sia nell'intervallo di confidenza è del* $100(1-\alpha)\%$.

Esempio 6.11 *Calcolare l'intervallo di confidenza* 95% *per la stima della media* μ *di una distribuzione normale con deviazione standard* $\sigma = 2$ *a partire da un campione di* 64 *elementi con media calcolata* $\bar{x} = 70$.

Soluzione.

Calcolo il quantile $z_{0.025} = 1.96$ della normale standard. Quindi l'intervallo di confidenza è:

$$\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} = 70 \pm 1.96 \cdot \frac{2}{\sqrt{(64)}} = 70 \pm 0.49.$$

Se considero un intervallo di confidenza 99% allora nella formula precedente devo modificare il valore del quantile: $z_{0.005} = 2.58$. L'intervallo diventa:

$$\bar{x} \pm 2.58 \cdot \frac{\sigma}{\sqrt{n}} = 70 \pm 2.58 \frac{2}{\sqrt{(64)}} = 70 \pm 0.645$$

quindi è più grande.

Sottolineiamo che l'intervallo di confidenza è un intervallo in cui si ha una certa *probabilità* di avere la stima esatta, quindi la sua interpretazione è ricondotta al concetto di probabilità di un evento.

Possiamo fare le seguenti considerazioni.

- Per un livello di confidenza fissato $1 - \alpha$, se n aumenta, l'intervallo di confidenza diminuisce.
- Per n fissato, se $1 - \alpha$ aumenta, l'intervallo di confidenza aumenta.

Esempio 6.12 *Calcolo intervallo di confidenza nel caso* $\alpha = 0.05$

Soluzione.

```
> sigma <- 2
> n <- 64
> mu <- 70
> alpha <- 0.05
> x <- rnorm(n, mu, sigma)
> mean(x)

## [1] 69.91232

> zalfa <- qnorm(1 - alpha / 2, 0, 1)
> mean(x) - zalfa * sigma / sqrt(n)

## [1] 69.42233

> mean(x) + zalfa * sigma / sqrt(n)

## [1] 70.40231
```

L' intervallo calcolato è relativo alla media campionaria $\bar{X}$:

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

quindi è un intervallo aleatorio, che cambia per ogni SRS(n) considerato.

Esempio 6.13 *Ripeto l'esempio precedente 100 volte e conto quante volte la media stimata sta nell'intervallo di confidenza.*

Soluzione.

```
> s <- 2
> alpha <- 0.05
> n <- 64
> k <- 0
> c1 <- vector(mode = "numeric")
> c2 <- vector(mode = "numeric")
> for (i in 1:100)
+ {
+   y <- rnorm(64, mean = 70, sd = 2)
+   m <- mean(y)
+   zalpha <- qnorm(1 - alpha / 2, mean = 0, sd = 1)
+   c1[i] <- m - zalpha * s / sqrt(n)
+   c2[i] <- m + zalpha * s / sqrt(n)
+   if (c1[i] <= 70 & 70 <= c2[i])
+   {
+     k <- k + 1
+   }
+ }
> k

## [1] 98
```

Il valore di k rappresenta il numero di volte che la media stimata sta nell'intervallo di confidenza in 100 SRS. Ricordiamo che in questo caso il livello di probabilità richiesto è 95%.

Problema. Se anche σ è non noto?

Si utilizza come intervallo di confidenza:

$$\bar{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}$$

dove S è la deviazione standard campionaria.

Cosa posso dire di questa approssimazione?

- Se n è grande, allora $\bar{X}$ ha una distribuzione che tende alla normale, quindi S sarà molto vicino al valore esatto σ.

Esempio 6.14 *Ripeto l'esempio precedente, sostituendo al posto della deviazione standard esatta σ quella approssimata S.*

```
> s <- 2
> alpha <- 0.05
> n <- 64
> k <- 0
> c1 <- vector(mode = "numeric")
> c2 <- vector(mode = "numeric")
> # se non ho la deviazione standard nota
> for (i in 1:100)
+ {
+   y <- rnorm(64, mean = 70, sd = s)
+   m <- mean(y)
+   zalpha <- qnorm(1 - alpha / 2, mean = 0, sd = 1)
+   S <- sd(y)
+   c1[i] <- m - zalpha * S / sqrt(n)
+   c2[i] <- m + zalpha * S / sqrt(n)
+   if (c1[i] <= 70 & 70 <= c2[i])
+   {
+     k <- k + 1
+   }
+ }
> k

## [1] 94
```

- se n è piccolo (e i campioni sono sempre estratti da una popolazione con distribuzione normale), allora posso sostituire il valore del quantile $z_{\alpha/2}$ con il quantile della distribuzione `t di student(df=1)`, perciò l'intervallo di confidenza diventa:

$$\bar{X} \pm t_{\alpha/2}(df = 1)\frac{S}{\sqrt{n}}.$$

Esempio 6.15 *Il tempo di risposta a un comando di una applicazione per tablet ha distribuzione normale con deviazione standard di 0.15 millisecondi. Si vuole stimare la media con un intervallo di confidenza 95% grande al più 10. Quale dimensione n del campione si deve utilizzare?*

Soluzione.

$$10 = 2 * (1.96)(15/\sqrt{n})$$

perciò:

$$n = (2 * 1.96 * 1.5)^2 = 34.57..$$

quindi $n \geq 35$.

6.2.5.2 Caso di popolazione di grande dimensione con distribuzione qualsiasi

Nel paragrafo precedente abbiamo considerato il caso di una popolazione con distribuzione normale. Se la popolazione invece NON ha distribuzione normale, cosa si pò dire?

Sia $X_1, X_2, \ldots X_n$ un SRS(n). Se n è sufficientemente grande per il *teorema del Limite Centrale* la variabile aleatoria $\bar{X}$ della media campionaria ha approssimativamente una distribuzione normale, qualunque sia la distribuzione di partenza. In particolare:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim norm(0, 1).$$

Tutte le considerazioni fatte nel paragrafo precedente per una distribuzione normale si possono quindi ripetere in questo caso sostituendo l'operatore di uguaglianza con un operatore di approssimazione per quanto riguarda la probabilità.

La difficoltà pratica di questo procedimento consiste nel fatto che nel caso di distribuzione non normale è veramente difficile conoscere la deviazione standard σ. Se si sostituisce a σ la devizione standard campionaria S si ha il seguente risultato:

Se n è sufficientemente grande, la variabile:

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \texttt{norm(0, 1)}$$

ha approssimativamente distribuzione normale standard, quindi l'intervallo:

$$\bar{x} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$$

è un intervallo di confidenza per μ con livello di confidenza *approssimativo* del $100(1-\alpha)\%$, INDIPENDENTEMENTE dalla distribuzione della popolazione dei campioni.

Esempio 6.16 *Facciamo una simulazione in R per verificare il risultato precedente.*

Soluzione.

```
> s <- 2
> alpha <- 0.05
> n <- 64
> k <- 0
> c1 <- vector(mode = "numeric")
> c2 <- vector(mode = "numeric")
> for (i in 1:100)
+ {
+   y <- rexp(64,rate=4)
+   m <- mean(y)
+   S <- sd(y)
+   zalpha <- qnorm(1 - alpha / 2, mean = 0, sd = 1)
+   c1[i] <- m - zalpha * S / sqrt(n)
+   c2[i] <- m + zalpha * S / sqrt(n)
+   if (c1[i] <= 1/4 & 1/4 <= c2[i])
+   {
+     k <- k + 1
+   }
+ }
> k

## [1] 90
```

In questo caso il valore di k è minore di 95, perché ho fatto un'approssimazione applicando il teorema del limite centrale. Rieseguendo il programma per esempio ottengo:

```
## [1] 95
```
