1 Metodo della massima verosimiglianza

Estraendo un campione costituito da n variabili casuali X_i i.i.d. da una popolazione X con funzione di probabilità/densità $f(x, \theta)$, si costruisce la FUNZIONE DI VEROSIMIGLIANZA[1] che rappresenta la funzione di probabilità/densità del campione stesso: in quest'ambito si ipotizza che essa sia funzione del vettore dei parametri θ, mentre le realizzazioni campionarie x_i sono fisse. Analiticamente si ha perciò:

$$L(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i, \theta) \qquad [1.1]$$

La funzione statistica $\hat{\theta} = t(x_1, x_2, \ldots, x_n)$ è detta STIMATORE DI MASSIMA VEROSIMIGLIANZA[2] se, in corrispondenza di ciascun campione estratto, assegna un valore al vettore θ che massimizza la funzione di verosimiglianza. In simboli:

$$\max L(x, \theta) = L(x, \hat{\theta}) \qquad [1.2]$$

Ovviamente la stima di massima verosimiglianza è definita in questo modo:

$$\hat{\theta} = \arg\max L(x, \theta) \qquad [1.3]$$

Per poter calcolare lo stimatore MLE si ricorre alla funzione LOG-VEROSIMIGLIANZA ottenuta attraverso l'applicazione del logaritmo naturale, quindi risulta:

$$\ell(x, \theta) = \ln L(x, \theta) \qquad [1.4]$$

Dato che la funzione logaritmica è una trasformazione monotòna crescente, con il passaggio alla log-verosimiglianza non si perdono le caratteristiche della funzione $L(x, \theta)$ in termini di crescenza e decrescenza e soprattutto si ottiene

[1] In inglese, *Likelihood function*.

[2] In letteratura tale stimatore è noto anche come "stimatore MLE", che deriva dall'inglese *Maximum Likelihood Estimator*. Per semplificare la notazione d'ora in avanti la funzione di verosimiglianza sarà indicata semplicemente con $L(x, \theta)$.

una forma analitica più semplice da trattare. Nel caso di variabili casuali i.i.d. questo è particolarmente vero perché la funzione di densità congiunta del campione può essere espressa come produttoria delle marginali: per le proprietà dei logaritmi, dalla [1.1] si ricava perciò la log-verosimiglianza come sommatoria, infatti:

$$\ell(x,\theta) = \ln\left[\prod_{i=1}^{n} f(x_i,\theta)\right] = \sum_{i=1}^{n} \ln f(x_i,\theta) \qquad [1.5]$$

Un'importante proprietà della log-verosimiglianza è la seguente:

$$\frac{\partial \ell(x,\theta)}{\partial \theta} = [L(x,\theta)]^{-1} \frac{\partial L(x,\theta)}{\partial \theta} \qquad [1.6]$$

Dimostrazione:

Applicando la derivata di una funzione logaritmica si ha:

$$\frac{\partial \ell(x,\theta)}{\partial \theta} = \frac{\partial \ln L(x,\theta)}{\partial \theta}$$
$$\frac{\partial \ell(x,\theta)}{\partial \theta} = \frac{\frac{\partial L(x,\theta)}{\partial \theta}}{L(x,\theta)}$$
$$\frac{\partial \ell(x,\theta)}{\partial \theta} = [L(x,\theta)]^{-1} \frac{\partial L(x,\theta)}{\partial \theta}$$

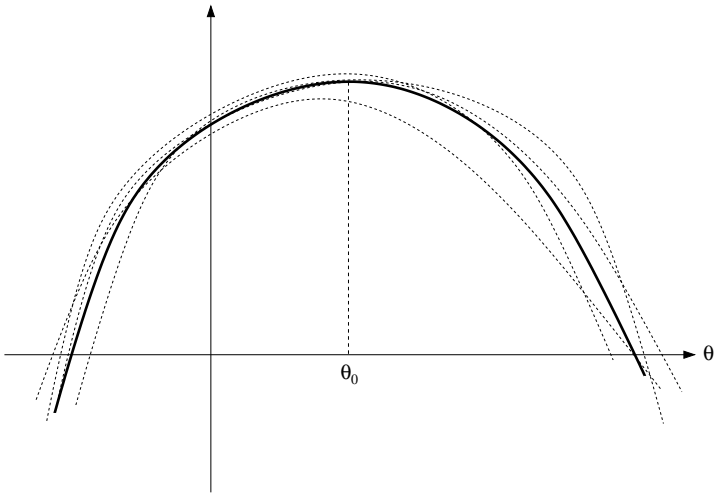
La più importante proprietà della funzione di log-verosimiglianza è però quella che costituisce la motivazione principale per il metodo di stima trattato: poiché nella sua definizione entrano le realizzazioni x_i delle n variabili casuali, la log-verosimiglianza è una funzione casuale del vettore dei parametri incogniti, nel senso che per θ dato, restituisce una variabile casuale; alternativamente si può pensare che ad ogni possibile realizzazione del campione è associata una diversa funzione di θ. Se questa funzione ha un valore atteso, tale valore atteso sarà una funzione non stocastica del vettore dei parametri.
Si osservi pertanto la figura 1.1: le linee tratteggiate corrispondono alla funzione di log-verosimiglianza osservabile per ciascuna realizzazione campionaria $f(x_1,\ldots,x_n;\theta_0)$, mentre la linea continua rappresenta la funzione $E[\ell(x,\theta)]$. Tale funzione assume un massimo proprio in corrispondenza di θ_0, ossia del vero valore della funzione di densità di x.

Dimostrazione:

Si consideri la seguente funzione di θ:

$$A(\theta) = E\left[\frac{L(x,\theta)}{L(x,\theta_0)}\right]$$

Figura 1.1 – Log-verosimiglianza campionaria e media



Questa funzione è uguale a 1 per $\forall\ \theta$, poiché $L(x,\theta)$ è una possibile funzione di densità di x per ciascun valore di θ, quindi:

$$A(\theta) = \int_x \frac{L(x,\theta)}{L(x,\theta_0)} L(x,\theta_0)dx = \int_x L(x,\theta)dx = 1$$

Da questa relazione segue che $\ln A(\theta) = 0$. Poiché il logaritmo è una funzione ovunque concava, per il lemma di Jensen si avrà:

$$\begin{aligned} &E\left[\ln \frac{L(x,\theta)}{L(x,\theta_0)}\right] \leq \ln E\left[\frac{L(x,\theta)}{L(x,\theta_0)}\right] = 0 \\ &E\left[\ell(x,\theta) - \ell(x,\theta_0)\right] \leq 0 \\ &E\left[\ell(x,\theta)\right] \leq E[\ell(x,\theta_0)] \end{aligned}$$

Tale relazione è vera per $\forall\ \theta \in \Theta$.

Alla luce di questo risultato è naturale pensare che il punto di massimo della funzione $\ell(x,\theta)$ possa essere utilizzato come stimatore di θ_0. È possibile dimostrare che, sotto condizioni molto generali, lo stimatore così ottenuto è uno stimatore consistente. Per la massimizzazione di $\ell(x,\theta)$ le condizioni del primo ordine risultano:

$$s(x,\theta) = \frac{\partial \ell(x,\theta)}{\partial \theta} = 0 \qquad [1.7]$$

dove la funzione $s(x,\theta)$ è detta SCORE. Quando si è in presenza di un campione composto da n variabili casuali i.i.d., lo score può anche essere scritto come segue:

$$s(x,\theta) = \sum_{i=1}^{n} s(x_i,\theta) = \sum_{i=1}^{n} \frac{\partial \ln f(x_i,\theta)}{\partial \theta} \qquad [1.8]$$

Dato che è funzione dei campioni estratti, lo score è anch'esso una variabile casuale; quando $\theta = \theta_0$, i suoi momenti sono:

1. $E[s(x, \theta_0)] = 0$

   Dimostrazione:

   Tenendo presente le proprietà delle derivate e la [1.6] risulta:

   $$E[s(x,\theta)] = \int_{-\infty}^{+\infty} s(x,\theta)L(x,\theta_0)dx$$
   $$E[s(x,\theta)] = \int_{-\infty}^{+\infty} \frac{\partial \ell(x,\theta)}{\partial \theta} L(x,\theta_0)dx$$
   $$E[s(x,\theta)] = \int_{-\infty}^{+\infty} \frac{\partial L(x,\theta)}{\partial \theta} [L(x,\theta)]^{-1} L(x,\theta_0)dx$$

   Valutando lo score per $\theta = \theta_0$, si ottiene:

   $$E[s(x,\theta_0)] = \int_{-\infty}^{+\infty} \frac{\partial L(x,\theta_0)}{\partial \theta} dx$$
   $$E[s(x,\theta_0)] = \frac{\partial}{\partial \theta} \int_{-\infty}^{+\infty} L(x,\theta_0)dx$$

   Poiché in corrispondenza del vero parametro θ_0 vale $\int_{-\infty}^{+\infty} L(x,\theta_0)dx = 1$, si ha:

   $$E[s(x,\theta_0)] = \frac{\partial 1}{\partial \theta} = 0$$

---

2. $Var[s(x, \theta_0)] = \mathcal{I}(\theta_0)$
   dove $\mathcal{I}(\theta_0)$ è la matrice di informazione di Fisher valutata in corrispondenza del vero parametro θ_0. La matrice $\mathcal{I}(\theta)$ è definita come l'opposto del valore atteso dell'Hessiana della log-verosimiglianza, quindi risulta:

   $$\mathcal{I}(\theta) = -E\left[\frac{\partial^2 \ell(x,\theta)}{\partial \theta \partial \theta'}\right] = -E\left[H(x,\theta)\right] \qquad [1.9]$$

   In corrispondenza di $\theta = \theta_0$ vale l'uguaglianza:

   $$\mathcal{I}(\theta_0) = Var[s(x,\theta_0)] = -E\left[H(x,\theta_0)\right] \qquad [1.10]$$

   Dimostrazione:

   Per dimostrare questa proprietà si parte dalla definizione di valore atteso dello score:

   $$E[s(x,\theta)] = \int_{-\infty}^{+\infty} s(x,\theta)L(x,\theta)dx$$

Derivando entrambi i membri rispetto a θ si ottiene:

$$\int_{-\infty}^{+\infty} \frac{\partial s(x,\theta) L(x,\theta)}{\partial \theta} dx = 0$$

Per la proprietà della derivata del prodotto e per la [1.6] si ha:

$$\begin{aligned}
&\int_{-\infty}^{+\infty} \frac{\partial s(x,\theta)}{\partial \theta} L(x,\theta) dx + \int_{-\infty}^{+\infty} \frac{\partial L(x,\theta)}{\partial \theta} s(x,\theta) dx = 0 \\
&\int_{-\infty}^{+\infty} \frac{\partial s(x,\theta)}{\partial \theta} L(x,\theta) dx + \int_{-\infty}^{+\infty} \frac{\partial \ell(x,\theta)}{\partial \theta} s(x,\theta) L(x,\theta) dx = 0 \\
&\int_{-\infty}^{+\infty} \frac{\partial s(x,\theta)}{\partial \theta} L(x,\theta) dx + \int_{-\infty}^{+\infty} s(x,\theta)^2 L(x,\theta) dx = 0
\end{aligned}$$

Utilizzando l'operatore valore atteso e valutando il tutto in $\theta = \theta_0$ si ha:

$$E\left[\frac{\partial s(x,\theta)}{\partial \theta}\right]_{\theta_0} + E\left[s(x,\theta_0)^2\right] = 0$$

Dato che $E[s(x,\theta_0)] = 0$ risulta:

$$\begin{aligned}
&E\left[\frac{\partial^2 \ell(x,\theta)}{\partial \theta \partial \theta'}\right]_{\theta_0} + Var[s(x,\theta_0)] = 0 \\
&Var[s(x,\theta)] = -E\left[\frac{\partial^2 \ell(x,\theta)}{\partial \theta \partial \theta'}\right]_{\theta_0} \\
&Var[s(x,\theta)] = -H[s(x,\theta_0)]
\end{aligned}$$

La matrice $\mathcal{I}(\theta_0)$ è posta uguale ad entrambi i membri di questa identità.

Nel caso di variabili casuali indipendenti la matrice di informazione di Fisher è data dalla sommatoria di tutte le varianze degli score che sono presenti all'interno della [1.8], infatti risulta:

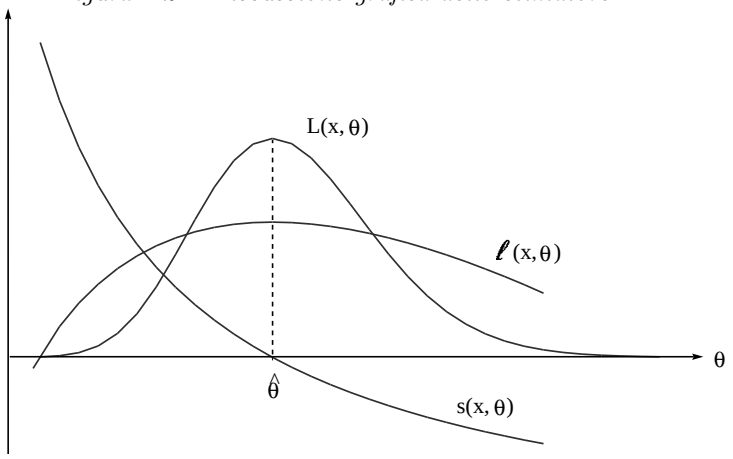
$$\mathcal{I}(\theta_0) = \sum_{i=1}^{n} Var[s(x_i, \theta_0)] = -\sum_{i=1}^{n} E\left[H(x_i, \theta_0)\right] \qquad [1.11]$$

Se le variabili casuali sono anche identicamente distribuite, tutti gli elementi della sommatoria sono uguali, per cui si può scrivere:

$$\mathcal{I}(\theta_0) = nVar[s(x_i, \theta_0)] = -nE\left[H(x_i, \theta_0)\right] \qquad [1.12]$$

La stima MLE è quindi quel valore del vettore θ in corrispondenza del quale lo score si annulla. La Figura 1.2 mostra graficamente come può essere ottenuta la stima di massima verosimiglianza.

Figura 1.2 – Discussione grafica dello stimatore MLE



All'interno del grafico sono state tracciate la funzione di verosimiglianza, la log-verosimiglianza e lo score[3]; in corrispondenza della stima MLE $\hat{\theta}$ si può agevolmente notare che le funzioni $L(x, \theta)$ e $\ell(x, \theta)$ hanno valore massimo, mentre $s(x, \theta)$ si annulla.

[3] Il grafico è stato tracciato per una variabile casuale di Poisson. Per fini esclusivamente didattici i valori della funzione di verosimiglianza sono stati moltiplicati per 10^6, mentre quelli della log-verosimiglianza sono stati traslati aggiungendo una costante pari a 25: in questo modo è possibile mettere a confronto le diverse curve all'interno di un solo grafico senza comprometterne le proprietà.

Esempio 1.1 *Calcolare lo stimatore MLE per l'incognito parametro π di una popolazione Bernoulliana con funzione di probabilità pari a:*

$$p(x,\pi) = \pi^x (1-\pi)^{1-x} \qquad \text{con } x_i = 0,1$$

Funzione di verosimiglianza:

$$L(x,\pi) = \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i} = \pi^{\gamma}(1-\pi)^{n-\gamma}$$

Passando alla log-verosimiglianza si ha:

$$\ell(x,\pi) = \ln \pi \sum_{i=1}^{n} x_i + \ln(1-\pi) \sum_{i=1}^{n}(1-x_i)$$
$$\ell(x,\pi) = \gamma \ln \pi + (n-\gamma)\ln(1-\pi)$$

dove $\gamma = \sum_{i=1}^{n} x_i$.

Derivando la log-verosimiglianza si ottiene lo score:

$$s(x,\pi) = \frac{\partial \ell(x,\pi)}{\partial \pi} = \frac{\gamma}{\pi} - \frac{n-\gamma}{1-\pi}$$

Eguagliando lo score a zero si ottiene lo stimatore di massima verosimiglianza.

$$\frac{\gamma}{\hat{\pi}} - \frac{n-\gamma}{1-\hat{\pi}} = 0$$
$$\gamma(1-\hat{\pi}) - \hat{\pi}(n-\gamma) = 0$$
$$\gamma - n\hat{\pi} = 0$$
$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Lo stimatore MLE per una popolazione Bernoulliana è la media campionaria.

Esempio 1.2 *Calcolare lo stimatore MLE per il parametro λ di una popolazione Poissoniana con la seguente funzione di probabilità:*

$$p(x,\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$$

Funzione di verosimiglianza:

$$L(x,\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\frac{\lambda^{\gamma}}{\prod_{i=1}^{n} x_i!}$$

dove $\gamma = \sum_{i=1}^{n} x_i$.
Funzione log-verosimiglianza:

$$\ell(x,\lambda) = -n\lambda + \ln\lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln(x_i!)$$

Funzione score:

$$s(x,\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i$$

Stimatore MLE:

$$-n + \frac{1}{\hat{\lambda}}\sum_{i=1}^{n} x_i = 0$$

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

Lo stimatore MLE per una popolazione Poissoniana coincide con la media campionaria.

Esempio 1.3 *Calcolare lo stimatore MLE per l'incognito vettore* $\theta = [\ \mu \quad \sigma^2\ ]'$ *di una popolazione normale. Si calcola la funzione di verosimiglianza:*

$$L(x,\theta) = \prod_{i=1}^{n}(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right\}$$

$$L(x,\theta) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\right\}$$

La log-verosimiglianza è:

$$\ell(x,\theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

lo score è il vettore gradiente contenente le derivate parziali della log-verosimiglianza calcolate rispetto ai parametri μ *e* σ^2*.*

$$s(x,\theta) = \begin{bmatrix} \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(x_i-\mu) \\ \\ -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(x_i-\mu)^2 \end{bmatrix}$$

Per ottenere lo stimatore MLE occorre risolvere il sistema:

$$\begin{bmatrix} \dfrac{1}{\hat{\sigma}^2}\displaystyle\sum_{i=1}^{n}(x_i-\hat{\mu}) \\ \\ -\dfrac{n}{2\hat{\sigma}^2} + \dfrac{1}{2\hat{\sigma}^4}\displaystyle\sum_{i=1}^{n}(x_i-\hat{\mu})^2 \end{bmatrix} = \begin{bmatrix} 0 \\ \\ 0 \end{bmatrix}$$

Dalla prima equazione si ottiene:

$$\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(x_i-\hat{\mu}) = 0$$

$$\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n} x_i = \frac{1}{\hat{\sigma}^2} n\hat{\mu}$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{X}$$

Sostituendo nella seconda si ha:

$$-\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

$$\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(x_i - \bar{X})^2 = n$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2 = s^2$$

Lo stimatore MLE per i parametri incogniti di una distribuzione normale è dato dalla media campionaria e dalla varianza campionaria. In simboli:

$$\hat{\theta} = \begin{bmatrix} \bar{X} \\ s^2 \end{bmatrix}$$

Estremo di Cramér-Rao

Esiste un valore minimo per la varianza di uno stimatore corretto? Se la risposta è affermativa significa che, se uno stimatore ha tale varianza, esso è sicuramente il più efficiente tra quelli corretti. Quando si è in presenza di un problema regolare di stima l'ESTREMO DI CRAMÉR-RAO fornisce la soluzione del quesito. Si ha un problema regolare di stima quando sono verificate le seguenti condizioni:

1. lo spazio dei parametri $\Theta \in \mathbb{R}^k$ è un insieme chiuso e limitato,

2. la funzione di probabilità/densità del campione $L(x_i, \theta)$ è continua e derivabile due volte rispetto al parametro θ,

3. il dominio della variabile casuale X non dipende da θ.

Per ricavare l'estremo di Cramér-Rao si considera un generico stimatore corretto per l'incognito parametro θ; per le proprietà dello score risulta:

$$Cov[\hat{\theta}, s(x, \theta_0)] = I_k \qquad [1.13]$$

dove I_k è la matrice identità di dimensione $k \times k$ e k è il numero di parametri contenuti in θ. L'importanza di questo risultato risiede nel fatto che la covarianza tra il generico stimatore corretto per θ e lo score valutato in θ_0 è costante a prescindere dal campione estratto.

Dimostrazione:

Partendo dall'equazione del valore atteso dello stimatore si ha:

$$\int_{-\infty}^{+\infty} \hat{\theta} L(x,\theta) dx = \theta$$

Derivando rispetto al vettore θ entrambi i membri (si tenga presente che lo stimatore è una funzione statistica quindi $\partial\hat{\theta}/\partial\theta = 0$) si ottiene:

$$\int_{-\infty}^{+\infty} \hat{\theta} \frac{\partial L(x,\theta)}{\partial\theta} dx = I_k$$

Ma poiché risulta:

$$\int_{-\infty}^{+\infty} \hat{\theta} \frac{\partial L(x,\theta)}{\partial\theta} dx = \int_{-\infty}^{+\infty} \hat{\theta} \frac{\partial \ell(x,\theta)}{\partial\theta} L(x,\theta) dx = \int_{-\infty}^{+\infty} \hat{\theta} s(x,\theta) L(x,\theta) dx = I_k,$$

Utilizzando l'operatore valore atteso si può perciò scrivere:

$$E[\hat{\theta} \cdot s(x,\theta)] = I_k$$

Valutando lo score in $\theta = \theta_0$ si ha $E[s(x,\theta_0)] = 0$, quindi, visto che $Cov(X,Y) = E(XY) - E(X)E(Y)$, si ottiene la seguente relazione:

$$Cov[\hat{\theta}, s(x,\theta_0)] = E[\hat{\theta} \cdot s(x,\theta_0)] = I_k$$

Dato il vettore $V = [\ \hat{\theta} \quad s(x,\theta_0)\ ]'$ risulta perciò:

$$Var(V) = \begin{bmatrix} Var(\hat{\theta}) & I_k \\ I_k & \mathcal{I}(\theta_0) \end{bmatrix} \qquad [1.14]$$

dove $Var(V)$ è una matrice almeno semidefinita positiva. Poiché risulta:

$$\begin{bmatrix} I_k & \mathcal{I}(\theta_0)^{-1} \end{bmatrix} \begin{bmatrix} Var(\hat{\theta}) & I_k \\ I_k & \mathcal{I}(\theta_0) \end{bmatrix} \begin{bmatrix} I_k \\ \mathcal{I}(\theta_0)^{-1} \end{bmatrix} = Var(\hat{\theta}) - \mathcal{I}(\theta_0)^{-1} \qquad [1.15]$$

Dato che anche la [1.15] è una matrice almeno semidefinita positiva[4], l'estremo di Cramér-Rao coincide con l'inversa della matrice di informazione di Fisher valutata per il vero parametro θ_0, quindi rappresenta il valore minimo per la varianza di uno stimatore corretto.

[4] Questa proprietà deriva dal seguente risultato: data una matrice A semidefinita positiva ed una qualsiasi matrice B, risulta semidefinita positiva anche la matrice BAB'.

Esempio 1.4 *Determinare l'estremo di Cramér-Rao per la media e la varianza di una popolazione* $X \sim N(\mu, \sigma^2)$.

La funzione di verosimiglianza, la log-verosimiglianza e la score sono state già determinate nell'Esempio 1.3. La derivata dello score rispetto ai parametri μ *e* σ^2 *è data dalla seguente matrice:*

$$\frac{\partial s(x,\theta)}{\partial \theta} = \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu) \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu) & \frac{n}{2\sigma^4} - \frac{2}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2 \end{bmatrix}$$

Calcolando il valore atteso e cambiando il segno si ottiene la matrice di informazione di Fisher.

$$\mathcal{I}(\theta) = -E\left[\frac{\partial s(x_i,\theta)}{\partial \theta}\right]$$

$$\mathcal{I}(\theta) = \begin{bmatrix} E\left(\frac{n}{\sigma^2}\right) & \frac{1}{\sigma^4}\sum_{i=1}^{n}E(x_i-\mu) \\ \frac{1}{\sigma^4}\sum_{i=1}^{n}E(x_i-\mu) & -\frac{n}{2\sigma^4} + \frac{2}{\sigma^6}\sum_{i=1}^{n}E(x_i-\mu)^2 \end{bmatrix} = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

L'estremo di Cramér-Rao è perciò:

$$\mathcal{I}(\theta)^{-1} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

PROPRIETÀ DELLO STIMATORE DI MASSIMA VEROSIMIGLIANZA:

1. Sotto alcune blande condizioni lo stimatore MLE è uno stimatore consistente in quanto risulta:
$$\hat{\theta} \xrightarrow{\Pr} \theta_0$$

2. Sotto opportune condizioni di regolarità, quasi sempre verificate nelle applicazioni pratiche, lo stimatore MLE ha la seguente distribuzione limite:
$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{\mathrm{d}} N\left(0, \Omega^{-1}\right)$$
dove $\Omega = (1/n)\mathcal{I}(\theta_0)$ è una matrice che non dipende dalla numerosità n del campione. Alla luce di questo risultato è possibile usare l'approssimazione:
$$\hat{\theta} \sim N\left(\theta_0, \mathcal{I}(\theta_0)^{-1}\right)$$
dove $\mathcal{I}(\theta_0)$ è la matrice di informazione di Fisher calcolata per il vero parametro θ_0.

3. Lo stimatore MLE risulta sempre asintoticamente efficiente. Questa proprietà deriva direttamente dalla precedente dato che, per $n \to \infty$, la varianza asintotica dello stimatore MLE coincide con l'estremo di Cramer-Rao dato dall'inversa della matrice di informazione di Fisher.

4. Lo stimatore MLE gode della proprietà di invarianza a trasformazioni monotòne continue: se il vettore θ segue una funzione $g(\theta)$ monotòna e continua, la stima di massima verosimiglianza vale $g(\hat{\theta})$.

Dimostrazione:

La funzione $g(\theta) : \Theta \longrightarrow \mathbb{R}$ trasforma ogni θ in un valore $k \in \mathbb{R}$, quindi risulta che:

$$\hat{k} = g(\hat{\theta})$$

Considerando la funzione di verosimiglianza si ha:

$$\begin{aligned} L(x,\theta) &= \prod_{i=1}^{n} f(x,\theta) \\ L(x,\theta) &= \prod_{i=1}^{n} f[x, g^{-1}(k)] \\ L(x,\theta) &= L[x, g^{-1}(k)] \\ L(x,\theta) &= L(x,k) \end{aligned}$$

Per le proprietà della funzione $g(\theta)$ la funzione di verosimiglianza è massimizzata per quel valore di k che si ottiene in corrispondenza dello stimatore MLE, infatti:

$$\begin{aligned} L(x,\hat{\theta}) &= L[x, g^{-1}(\hat{k})] \\ L(x,\hat{\theta}) &= L(x,\hat{k}) \end{aligned}$$

Quest'ultima proprietà è molto importante soprattutto per due ragioni:

a) Se si conosce il valore di $\hat{\theta}$ e si desidera stimare col metodo della massima verosimiglianza una funzione del vero parametro incognito, non occorre ripetere il procedimento di stima.

b) È possibile applicare qualsiasi trasformazione monotòna continua alla funzione $L(x,\theta)$ in modo da renderla più semplice da gestire a livello analitico, senza implicare mutamenti nei valori assunti dallo stimatore MLE.

2 Test basati sulla verosimiglianza

Una statistica test è una statistica che ha una distribuzione nota (almeno per $n \to \infty$) sotto l'ipotesi nulla, cosicché è possibile costruire una regione di accettazione e una regione di rifiuto. Ma come si costruisce una statistica test? Uno degli approcci più usati in econometria è basato sulla verosimiglianza, ed è applicabile in tutte le situazioni in cui l'ipotesi nulla possa essere scritta nella forma:

$$g(\theta_0) = 0$$

dove la funzione $g(\theta_0)$ è continua, derivabile, ha come dominio lo spazio dei parametri $\Theta \in \mathbb{R}^k$ e come codominio $\mathbb{R}^q$. In altri termini, questa funzione definisce un sistema di q vincoli sui k parametri oggetto di stima. Naturalmente, si ha che $q \leq k$.

Esempio 2.1 *Dato un campione di variabili casuali normali con media μ e varianza σ^2, il vettore dei parametri incogniti θ è:*

$$\theta = [\theta_1 \quad \theta_2]' = \begin{bmatrix}\mu & \sigma^2\end{bmatrix}'$$

La funzione $g(\theta)$ corrispondente all'ipotesi $H_0 : \mu = 0$ è semplicemente:

$$g(\theta) = \theta_1$$

Un'ipotesi nulla più articolata, come ad esempio $H_0 : \mu = \sigma$ avrebbe potuto essere espressa per mezzo della funzione

$$g(\theta) = \theta_1 - \sqrt{\theta_2}$$

oppure indifferentemente,

$$g(\theta) = \frac{\theta_1}{\sqrt{\theta_2}} - 1$$

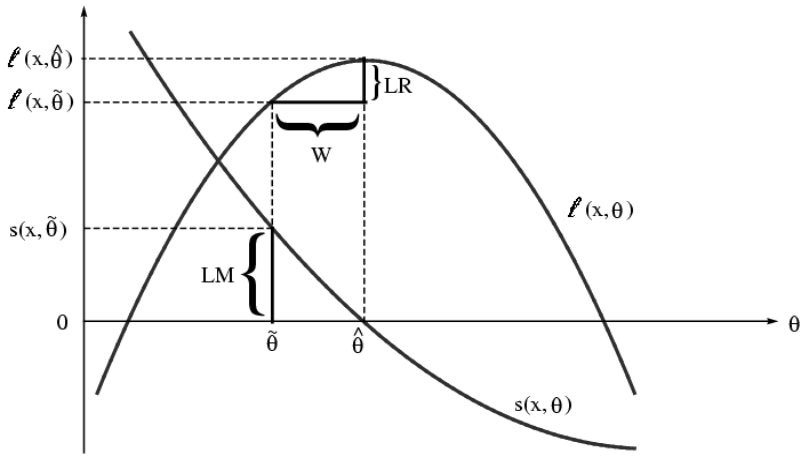
La stima di massima verosimiglianza può essere effettuata tenendo conto o meno dei vincoli $g(\theta_0) = 0$. Lo stimatore ottenuto senza tener conto dei vincoli (già considerato nel sottoparagrafo 1, e generalmente indicato con la

notazione $\hat{\theta}$) si chiama stimatore LIBERO. Si consideri il seguente problema di ottimizzazione:

$$\begin{cases} \text{Max} & L(x,\theta) \\ \text{Sub} & g(\theta) = 0 \end{cases} \qquad [2.1]$$

Lo stimatore ottenuto come soluzione del problema, prende il nome di stimatore VINCOLATO, e di solito si indica con il simbolo $\tilde{\theta}$. Si consideri la Figura 2.1:

Figura 2.1 – I tre test classici



l'idea di base è quella di confrontare la soluzione libera con quella vincolata e considerare valida l'ipotesi nulla se le due soluzioni non sono "troppo distanti". In pratica, questo può essere fatto in tre modi diversi, ciò che conduce a definire tre tipi di test:

Test LR: la sigla LR sta per *Likelihood Ratio*, in italiano "rapporto di verosimiglianza"; questo tipo di test è basato sul confronto fra il massimo libero e il massimo vincolato, ossia sul confronto fra $\ell(x,\hat{\theta})$ e $\ell(x,\tilde{\theta})$.

Test W: noto come *Test di Wald*, esso pone direttamente a confronto gli stimatori libero $\hat{\theta}$ e vincolato $\tilde{\theta}$, anziché i rispettivi valori della verosimiglianza.

Test LM: anche detto "test dello *score*", si basa sul valore dello score in $\tilde{\theta}$ (ovviamente lo score valutato in $\hat{\theta}$ è nullo per costruzione). La sigla LM sta per *Lagrange Multipliers*.

Osservando nuovamente la Figura 2.1 si può notare che il test LR è basato sulla differenza in ordinata fra la soluzione libera e quella vincolata, mentre il test W considera la differenza in ascissa. Il test LM, invece, è basato sulla pendenza della log-verosimiglianza nel punto di massimo vincolato, ossia sul

valore dello score in $\tilde{\theta}$.
Un aspetto importante è che questi tre test sono asintoticamente equivalenti[1], cioè, sotto alcune blande condizioni di regolarità, risulta che:

$$\text{plim}(LR - W) = \text{plim}(LR - LM) = \text{plim}(W - LM) = 0 \qquad [2.2]$$

Da questa proprietà deriva che essi hanno anche la stessa distribuzione asintotica, che solitamente risulta essere una variabile casuale χ^2_q, dove il numero dei g.d.l. q è pari a quello dei vincoli presenti all'interno della funzione $g(\theta)$. Il test relativo all'ipotesi nulla è pertanto un test ad una coda, con regione di accettazione pari a $[0, c_\alpha]$, dove α è il livello di significatività e c_α è il relativo valore critico della χ^2_q. L'esempio 2.5 rappresenta un caso particolare di questa proprietà.
Per popolazioni distribuite normalmente, quando si è in presenza di campioni di numerosità finita, vale inoltre la relazione:

$$W \geq LR \geq LM \qquad [2.3]$$

2.0.1 Test LR

Il test LR è il più semplice da calcolare, una volta che le soluzioni dei problemi di ottimo libero e vincolato siano entrambe disponibili. La forma della statistica test è la seguente:

$$LR = 2\left[\ell(x, \hat{\theta}) - \ell(x, \tilde{\theta})\right] \qquad [2.4]$$

Il nome è dovuto al fatto che la differenza fra le log-verosimiglianze è ovviamente uguale al logaritmo naturale del rapporto fra le funzioni di verosimiglianza, cioè:

$$LR = 2\ln\left[\frac{L(x, \hat{\theta})}{L(x, \tilde{\theta})}\right] \qquad [2.5]$$

Esempio 2.2 *Sia dato un campione di variabili X_i, distribuite come esponenziali negative i.i.d. con le seguenti caratteristiche: ampiezza campionaria $n = 400$; media aritmetica $\bar{X} = 0.8$. Calcolare un test LR per l'ipotesi $H_0 : \theta_0 = 1$.*

Poiché la funzione di densità dell'i-esima osservazione è:

$$f(x_i, \theta) = \theta e^{-\theta x_i}$$

La funzione di log-verosimiglianza è data da

$$\ell(\theta) = n(\ln\theta - \theta\bar{X})$$

Si deduce facilmente che la stima libera di θ è pari a:

$$\hat{\theta} = \frac{1}{\bar{X}} = 1.25$$

[1] Si veda in proposito l'Esempio 2.5.

Il valore della log-verosimiglianza nel punto di massimo è pari a:

$$\ell(\hat{\theta}) = n\,(\ln 1.25 - 1) = 400 \cdot (-0.7769) = -310.7426$$

Il punto di massimo vincolato è l'unico valore possibile per θ sotto l'ipotesi nulla, ossia $\tilde{\theta} = \theta_0 = 1$; la log-verosimiglianza corrispondente è

$$\ell(\tilde{\theta}) = n(\ln 1 - 1 \cdot 0.8) = 400 \cdot (-0.8) = -320$$

Applicando la [2.4] risulta:

$$LR = 2\,[-310.7426 - (-320)] = 18.515$$

Sotto l'ipotesi nulla, il test si distribuisce come una χ^2_1. I valori critici per tale distribuzione sono:

α	χ^2_1
10%	2.7055
5%	3.8414
1%	6.6349

L'ipotesi viene rifiutata a qualunque livello di significatività.

2.0.2 Test W

Il test di Wald è reso possibile dalla normalità asintotica dello stimatore di massima verosimiglianza[2], espressa dalla seguente espressione:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{\text{d}} N\left(0, A_n\right)$$

dove $A_n = n\mathcal{I}(\theta_0)^{-1}$. Se questo è vero, è possibile utilizzare i risultati di teoria asintotica (in particolare il "delta method") per affermare che, sotto l'ipotesi nulla, vale la relazione:

$$\sqrt{n}g(\hat{\theta}) \xrightarrow{\text{d}} N\left(0, \Omega\right) \qquad [2.6]$$

con $\Omega = GA_nG'$ e $G = \left.\dfrac{\partial g(\theta)}{\partial \theta}\right|_{\theta_0}$.
In pratica, si parte dalla distribuzione asintotica dello stimatore per calcolare la distribuzione asintotica della funzione vincolo $g(\hat{\theta})$. Pertanto, se H_0 è vera, la statistica test di Wald è data dalla seguente forma quadratica:

$$W = n\left[g(\hat{\theta})'\Omega^{-1}g(\hat{\theta})\right] \qquad [2.7]$$

La [2.7] converge ad una variabile casuale χ^2_q dove q rappresenta il numero di vincoli imposti sui parametri dalla funzione $g(\hat{\theta})$. Nel caso in cui la matrice Ω non fosse osservabile, è sufficiente dal punto di vista asintotico usare un suo stimatore consistente.

[2]Si veda in proposito la proprietà 2. di pag. 11.

Esempio 2.3 *Con gli stessi dati dell'esempio 2.2, calcolare un test di Wald.*
Distribuzione asintotica dello stimatore $\hat{\theta}$: derivando due volte la funzione di log-verosimiglianza si ottiene:

$$H(\theta) = \frac{\partial^2 \ell(\theta)}{\partial \theta^2} = \frac{\partial}{\partial \theta}\left[n\left(\frac{1}{\theta} - \bar{X}\right)\right] = -\frac{n}{\theta^2}$$

La matrice di informazione di Fisher è pari al negativo del valore atteso dell'Hessiano quindi risulta essere uguale a:

$$\mathcal{I}(\theta) = \frac{n}{\theta^2}$$

quindi

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{\text{d}} N\left(0, \theta^2\right)$$

Il vincolo in questo caso è molto semplice in quanto si ha $g(\theta) = \theta - 1$ e di conseguenza $G = 1$.
Il test di Wald, pertanto può basarsi sul fatto che, sotto l'ipotesi nulla, risulta:

$$\sqrt{n}\left(\hat{\theta} - 1\right) \xrightarrow{\text{d}} N\left(0, 1\right)$$

La statistica test ha quindi la seguente distribuzione:

$$W = n\left(\hat{\theta} - 1\right)^2 \sim \chi^2_1$$

Con i dati dell'esempio si ha:

$$W = 400 \cdot (1.25 - 1)^2 = 400 \cdot 0.0625 = 25$$

Anche in questo caso l'ipotesi nulla viene rifiutata.

2.0.3 Test LM

Il criterio su cui si basa il test LM è il seguente: se l'ipotesi nulla è vera, allora lo stimatore vincolato $\tilde{\theta}$ deve essere consistente. In questo caso la funzione casuale $s(x, \theta)$ asintoticamente deve avere le stesse proprietà sia che venga calcolata nel "vero" parametro θ_0, sia in $\tilde{\theta}$. Nell'ipotesi di variabili i.i.d. risulta che:

- $E\left[s(x_i, \theta_0)\right] = 0$
- $Var\left[s(x_i, \theta_0)\right] = \overline{\mathcal{I}}(\theta_0)$

dove $n\overline{\mathcal{I}}(\theta_0) = \mathcal{I}(\theta_0)$. A questo punto è possibile invocare il TLC di Lindeberg-Lévy per affermare la normalità asintotica della media aritmetica degli score, cioè:

$$\frac{1}{\sqrt{n}} s(x, \theta_0) = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} s(x_i, \theta_0)\right] \xrightarrow{\text{d}} N\left(0, \overline{\mathcal{I}}(\theta_0)\right) \qquad [2.8]$$

Costruendo la forma quadratica si ha:

$$\frac{1}{n}s(x,\theta_0)'\overline{\mathcal{I}}(\theta_0)^{-1}s(x,\theta_0) = s(x,\theta_0)'\mathcal{I}(\theta_0)^{-1}s(x,\theta_0) \xrightarrow{\text{d}} \chi^2_q \qquad [2.9]$$

Poiché se H_0 è vera risulta che $\tilde{\theta} \xrightarrow{\text{Pr}} \theta_0$, allora ne consegue anche che:

$$LM = s(x,\tilde{\theta})'\mathcal{I}(\tilde{\theta})^{-1}s(x,\tilde{\theta}) \xrightarrow{\text{d}} \chi^2_q \qquad [2.10]$$

È possibile dimostrare che questo test è asintoticamente equivalente ai test LR e W, cosicché il numero di g.d.l. del test è pari al numero dei vincoli q. Come risultato, è possibile impostare un test a una coda la cui regione di rifiuto è data da:

$$LM > c_\alpha$$

dove c_α è quel valore tale per cui $Pr(\chi^2_q > c_\alpha) = \alpha$.

Esempio 2.4 *Con gli stessi dati dell'esempio 2.2, calcolare un test LM.*

Nell'esempio 2.3 abbiamo già calcolato score e matrice di informazione, per cui non resta che da applicare la [2.10].

$$s(\tilde{\theta}) = n\left(\frac{1}{\tilde{\theta}} - \bar{X}\right) = n(1-\bar{X})$$
$$\mathcal{I}(\tilde{\theta}) = \mathcal{I}(1) = n$$

La statistica test è:

$$LM = n(1-\bar{X})n^{-1}n(1-\bar{X}) = n(1-\bar{X})^2$$

Essa ha una distribuzione asintotica χ^2 con 1 g.d.l. Si noti che in questo particolare contesto il test LM è identico al test W quindi la stessa statistica test può essere interpretata in entrambi i modi. Come nell'esempio 2.3 risulta:

$$LM = 400 \cdot (1-0.8)^2 = 400 \cdot 0.004 = 16$$

Anche in quest'ambito l'ipotesi nulla non può essere accettata. Si noti inoltre che la condizione [2.3] è rispettata.

2.0.4 Criteri di scelta fra i tre test classici

Dovrebbe essere chiaro dalla discussione precedente che la scelta di quale dei tre test classici usare in pratica non può essere basato su considerazioni di natura asintotica, visto che i test sono asintoticamente equivalenti. In certi contesti è possibile impostare un confronto sulle proprietà dei test in campioni finiti, ma di regola questa è una via difficilmente praticabile.
Il criterio che si utilizza di solito è quello della semplicità di calcolo: infatti,

ognuno dei tre test ha la particolarità di richiedere, per il suo calcolo, informazioni che non sono richieste per gli altri due.
La differenza è evidente fra i test LM e W: infatti, il primo è calcolabile solo a partire dal punto di massimo vincolato, mentre il secondo richede la sola conoscenza del massimo libero (si vedano gli esempi 2.3 e 2.4). Il test LR viceversa, richiede la conoscenza di ambo le soluzioni. Esso, tuttavia, non richiede il calcolo delle derivate della funzione di log-verosimiglianza (necessario negli altri due casi), in quanto è sufficiente calcolare la differenza fra $\ell(x, \hat{\theta})$ e $\ell(x, \tilde{\theta})$ e moltiplicarla per due.

Esempio 2.5 *Data una popolazione $X \sim N(\mu, \sigma^2/n)$ con σ^2 noto, data $H_0 : \mu = \mu_0$, si verifichi che i test LR, W e LM coincidono per $n \to \infty$.*

1. *Funzione log-verosimiglianza:*

$$\ell(x, \mu) = k - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$$

2. *Score:*

$$s(x, \mu) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)$$

3. *Inversa della matrice di informazione di Fisher:*

$$\mathcal{I}(\mu)^{-1} = \frac{\sigma^2}{n}$$

$$LR = 2 \left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \hat{\mu})^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu_0)^2 \right]$$

$$LR = \frac{1}{\sigma^2} \left[-\sum_{i=1}^{n} (X_i - \hat{\mu})^2 + \sum_{i=1}^{n} (X_i - \mu_0)^2 \right]$$

$$LR = \frac{n}{\sigma^2} \left(\bar{X} - \mu_0 \right)^2$$

$$W = \frac{(\hat{\mu} - \mu_0)^2}{\mathcal{I}(\hat{\mu})}$$

$$W = \frac{n}{\sigma^2} (\bar{X} - \mu_0)^2$$

$$LM = \left[\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu_0) \right]^2 \frac{\sigma^2}{n}$$

$$LM = \frac{1}{n\sigma^2} \left[n\bar{X} - n\mu_0 \right]^2$$

$$LM = \frac{n}{\sigma^2} (\bar{X} - \mu_0)^2$$

Il risultato a cui si giunge è LR=W=LM=$\frac{n}{\sigma^2}(\bar{X} - \mu_0)^2$. Quando $n \to \infty$ si ha perciò l'equivalenza asintotica dei test classici, cioè:

$$LR,\ W,\ LM \sim \chi^2_1$$

Esempio 2.6 *Negli esempi 1.3 e 1.4 sono state determinate le equazioni dello stimatore ML, della log-verosimiglianza e della matrice di informazione di Fisher relative ad una popolazione distribuita come una variabile casuale normale. Utilizzando tali informazioni calcolare i test LR, W e LM relativi all'ipotesi nulla che la popolazione abbia distribuzione normale standardizzata, sapendo che:*

- $\sum_{i=1}^{n} X_i = 100$
- $\sum_{i=1}^{n} X_i^2 = 470$
- $n = 500$

L'ipotesi nulla è:

$$H_0 : \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

o alternativamente

$$\begin{cases} \tilde{\mu} = 0 \\ \tilde{\sigma}^2 = 1 \end{cases}$$

Gli stimatori ML per la media e per la varianza sono rispettivamente:

- $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = 0.2$
- $s^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = 0.9$

1. *Test LR*

$$LR = 2 \left[-\frac{n}{2} \ln \hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2} \sum_{i=1}^{n} (X_i - \hat{\mu})^2 + \frac{n}{2} \ln \tilde{\sigma}^2 + \frac{1}{2\tilde{\sigma}^2} \sum_{i=1}^{n} (X_i - \tilde{\mu})^2 \right]$$
$$LR = \left[-n \ln \hat{\sigma}^2 - n + \sum_{i=1}^{n} X_i^2 \right]$$
$$LR = [-500 \cdot (-0.105) - 500 + 470] = 22.5$$

2. *Test W*
   La matrice Jacobiana calcolata sul vincolo è $G = I$ dove I è la matrice identità. La matrice Ω^{-1} sarà perciò:

$$\Omega^{-1} = n\mathcal{I}(\hat{\mu}, \hat{\sigma}^2) = \begin{bmatrix} \hat{\sigma}^2 & 0 \\ 0 & 2\hat{\sigma}^2 \end{bmatrix}$$

   Si costruisce pertanto la seguente forma quadratica:

$$W = n \begin{bmatrix} \hat{\mu} & \hat{\sigma}^2 - 1 \end{bmatrix} \begin{bmatrix} \hat{\sigma}^2 & 0 \\ 0 & 2\hat{\sigma}^4 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 - 1 \end{bmatrix}$$
$$W = 500 \begin{bmatrix} 0.2 & -0.1 \end{bmatrix} \begin{bmatrix} 0.9 & 0 \\ 0 & 1.62 \end{bmatrix} \begin{bmatrix} 0.2 \\ -0.1 \end{bmatrix} = 26.1$$

3. *Test LM*

   Lo score nel modello vincolato é:

$$s(x, \tilde{\mu}, \tilde{\sigma}^2) = \begin{bmatrix} \frac{1}{\tilde{\sigma}^2} \sum_{i=1}^{n} (X_i - \tilde{\mu}) \\ -\frac{n}{2\tilde{\sigma}^2} + \frac{1}{2\tilde{\sigma}^4} \sum_{i=1}^{n} (X_i - \tilde{\mu})^2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} X_i \\ -\frac{n}{2} + \frac{1}{2} \sum_{i=1}^{n} X_i^2 \end{bmatrix} = \begin{bmatrix} 100 \\ 25 \end{bmatrix}$$

   L'inversa della matrice di informazione di Fisher calcolata sotto H_0 è:

$$\mathcal{I}(\tilde{\mu}, \tilde{\sigma}^2) = \begin{bmatrix} \frac{\tilde{\sigma}^2}{n} & 0 \\ 0 & \frac{2\tilde{\sigma}^4}{n} \end{bmatrix} = \begin{bmatrix} \frac{1}{500} & 0 \\ 0 & \frac{2}{500} \end{bmatrix}$$

   La statisica test è perciò:

$$LM = \begin{bmatrix} 100 & 25 \end{bmatrix} \begin{bmatrix} \frac{1}{500} & 0 \\ 0 & \frac{2}{500} \end{bmatrix} \begin{bmatrix} 100 \\ 25 \end{bmatrix} = 22.5$$

Poiché $\chi^2_2 = 5.9914$ l'ipotesi nulla non può essere accettata.
