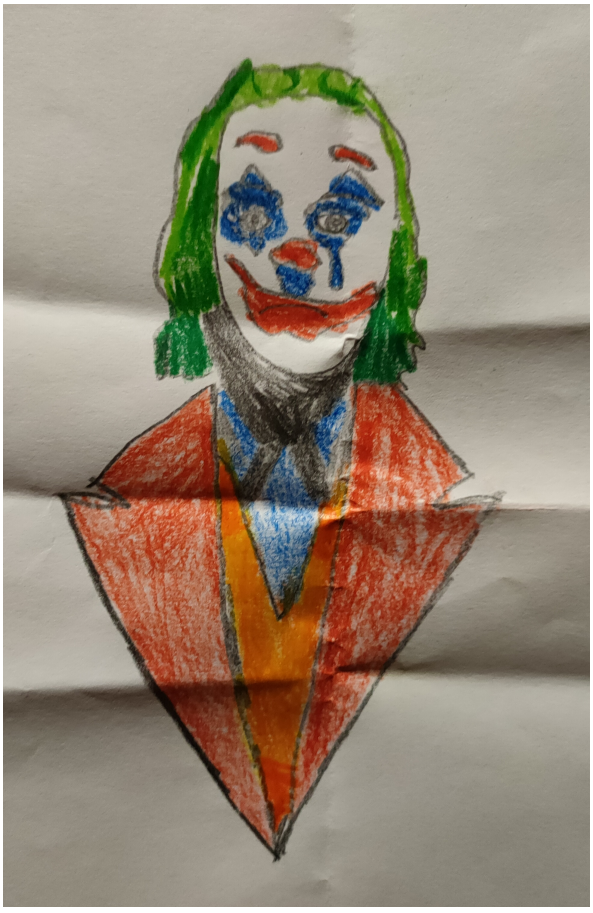
Gianfranco Adimari

Appunti di Inferenza: la verosimiglianza



Premessa

Queste pagine presentano una parte degli argomenti proposti nell'insegnamento *Statistica 2* - insegnamento obbligatorio dei corsi di laurea triennale offerti dal Dipartimento di Scienze Statistiche di Padova-. Più precisamente, gli "Appunti" trattano, ad un livello introduttivo, le tecniche di inferenza statistica basate sulla funzione di verosimiglianza. Inoltre, nella parte finale, propongono alcuni esercizi riepilogativi, che si riferiscono agli argomenti trattati nell'intero Corso.

Lo scopo degli "Appunti" è quello di fornire al lettore uno strumento che lo aiuti a seguire le lezioni e comprendere gli elementi di teoria della Statistica basata sulla verosimiglianza. Naturalmente, al lettore è richiesta la conoscenza delle nozioni trattate nella prima parte dell'insegnamento (introduzione ai problemi di stima puntuale, intervallare e di verifica d'ipotesi, con applicazione a casi specifici) oltre che delle nozioni trattate nei corsi di base di Analisi Matematica e Calcolo delle Probabilità.

La stesura degli "Appunti" prende spunto da vario materiale ereditato dalle colleghe Monica Chiogna e Alessandra Salvan, che ringrazio.

Come testi teorici di riferimento si segnalano, in particolare, quelli suggeriti per il Corso:

- *Introduzione alla Statistica. II Inferenza, Verosimiglianza, Modelli*, di Luigi Pace e Alessandra Salvan (CEDAM, Padova, 2001);
- *Inferenza Statistica. Una presentazione basata sul concetto di verosimiglianza*, di Adelchi Azzalini (Springer, 2000).

Padova, 6 gennaio 2024 G.A.

0.1 Verso la funzione di verosimiglianza

Sebbene in alcune situazioni (tipicamente più semplici) il metodo dell'analogia possa costituire un'importante risorsa per il reperimento di adeguati strumenti inferenziali, quali stimatori, quantità pivotali o statistiche test, in moltissimi altri contesti, connotati da crescente complicazione, l'approccio basato su tale metodo risulta praticamente inutile. L'esempio che segue aiuta a chiarire.

Esempio: dati relativi ai voli della navicella Shuttle.
L'esito della missione spaziale USA Challenger del 27 gennaio 1986 fu catastrofico. La navicella esplose subito dopo il lancio. Analisi successive mostrarono che l'esplosione fu causata dalla mancata tenuta di alcune guarnizioni (O-rings) dei motori. Si sospettò che la tenuta delle guarnizioni potesse dipendere dalla temperatura esterna al momento del lancio, che quel giorno era pari a $31°$ F (circa $-0,6°$ C).

volo n.	y_i	temperatura ° F, t_i
1	0	66
2	1	70
3	0	69
4	0	68
5	0	67
6	0	72
7	0	73
8	0	70
9	1	57
10	1	63
11	1	70
12	0	78
13	0	67
14	1	53
15	0	67
16	0	75
17	0	70
18	0	81
19	0	76
20	0	79
21	1	75
22	0	76
23	1	58

La tabella a fianco riporta i dati relativi a 23 lanci di prova (o simulazioni di volo), effettuate precedentemente al volo di quel giorno. I valori y_i valgono 1 o 0 a seconda che si sia registrato almeno una rottura per le guarnizioni o meno. I valori t_i sono quelli della temperatura esterna presente al momento della prova (o della simulazione). Questi dati possono essere utilizzati per fornire risposte a due domande: (i) la probabilità di rottura di almeno una guarnizione dipende (in maniera significativa - statisticamente) dalla temperatura esterna? (ii) qual è una stima ragionevole della probabilità di rottura il giorno del disastro? Naturalmente, per procedere con questo obiettivo è necessario fissare delle *ipotesi di lavoro*, a cui rimarranno condizionati i risultati che si otterranno.

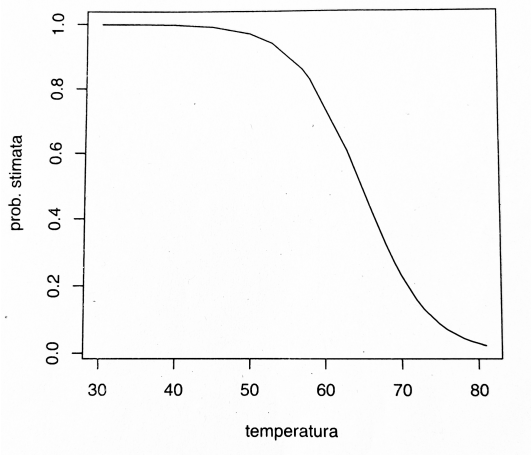
Nel caso specifico, vista la natura dei valori y_i, è del tutto naturale assumere che essi siano realizzazioni di variabili, diciamo Y_i, di Bernoulli. Inoltre per poter rispondere alla prima domanda è necessario assumere che i parametri π_i, caratterizzanti le leggi di tali variabili, possano (in generale) variare con la temperatura, e quindi siano funzioni dei valori t_i, cioè, assumiamo che sia $\pi_i = E(Y_i) = g(ti)$ per una qualche funzione g da scegliere opportunamente. In questa analisi i valori t_i li pensiamo fissati, quindi non casuali. Si usa dire, in casi come questo, che si lavora *condizionatamente* ai valori t_i. Se assumiamo l'indipendenza tra le 23 prove, rimane da scegiere la forma di g. Volendo semplificarci la vita, potremmo pensare alla funzione lineare $\alpha + \beta t_i$, che rappresenta la relazione più semplice. In realtà, dovendo modellare una risposta (il valore atteso di Y_i) che assume valori su (0,1), dovremmo imporre ai parametri α e β vincoli adeguati a garantire che il modello rispetti la natura della risposta. In casi come questo è più conveniente trasformare preventivamente la risposta, mediante una trasformazione (biunivoca) che ci liberi dal problema dei vincoli sui parametri del modellino lineare. Una trasformazione molto utilizzata per risposte a valori su (0,1) è la trasformazione *logit*: $logit(\pi_i) = \log(\pi_i/(1-\pi_i))$. In definitiva arriviamo ad assumere che le variabili Y_i sono variabili di Bernoulli indipendenti, per le quali vale la relazione $logit(E(Y_i)) = \alpha + \beta t_i$. È importante notare due cose. La prima è che questa formulazione

contempla solo due entità (parametri) ignote, α e β; la seconda è che, se decidiamo di proseguire con qualunque operazione di inferenza, accettiamo implicitamente il modello formalizzato come adeguata sintesi della realtà.

Ciò detto, come possiamo fare inferenza sui parametri ignoti del modello fissato, o su qualunque altra entità di interesse, disponendo dell'osservazione costituita dalle coppie (y_i, t_i), $i = 1, \ldots, 23$? In particolare come possiamo pensare di stimare α e β? Ci può per caso venire in aiuto il metodo dell'analogia? La risposta è no; perche in questo caso i parametri ignoti non rappresentano particolari caratteristiche della legge che genera i dati di cui possiamo trovare analoghe quantità campionarie (come avviene quando l'interesse riguarda una media o una varianza o una proporzione). Fortunatamente, esiste uno strumento molto generale per l'inferenza che è la **funzione di verosimiglianza**, di cui tratteremo ampiamente in queste pagine, e che consente di risolvere problemi di stima puntuale e ricavare tecniche per la costruzione di intervalli (o regioni) di confidenza e per la soluzione di problemi di verifica d'ipotesi, in situazioni come quella che stiamo trattando in questo esempio e anche molto più complesse.
Con riferimento alle due domande (i) e (ii) poste sopra, è chiaro che, una volta accettato che il modello scelto possa ben descrivere la realtà, per rispondere alla domanda (i) bisogna risolvere un problema di verifica d'ipotesi in cui è $H_0 : \beta = 0$. Per quanto riguarda la domanda (ii), una volta ottenute stime adeguate $\hat{\alpha}$ e $\hat{\beta}$, la stima della probabilità di rottura alla temperatura esterna di $31°$ F, sarà data da

$$\hat{\pi}_* = \frac{e^{\hat{\alpha}+\hat{\beta}t^*}}{1 + e^{\hat{\alpha}+\hat{\beta}t^*}},$$

con $t^* = 31$.



L'approccio basato sulla funzione di verosimiglianza porta ad ottenere le stime $\hat{\alpha} = 15,043$ e $\hat{\beta} = -0,2322$. La figura a fianco mostra l'andamento della stima della probabilià di rottura di almeno una guarnizione, al variare della temperatura, secondo il modello scelto. È chiaro il livello di preoccupazione che avrebbe dovuto creare la temperatura rilevata il giorno del disastro!

0.2 La funzione di verosimiglianza

La verosimiglianza è uno strumento parametrico. Questo significa che per poter definire la funzione di verosimiglianza è necessario scegliere, in via preliminare, un **modello statistico parametrico** per la variabile casuale da cui si pensa generata l'intera osservazione campionaria. Un modello statistico parametrico $\mathcal{F}$ è una classe di leggi, tutte definite sullo stesso supporto, indicizzate da un parametro ignoto, diciamo θ. Al variare di θ in un insieme di valori possibili che chiameremo **spazio parametrico** e indicheremo spesso con Θ, vengono individuate tutte le leggi della classe.

Θ deve essere un sottoinsieme (o al più coincidere) dello (con lo) spazio euclideo di dimensione finita, diciamo k; formalmente, $\Theta \subseteq \Re^k$. Inoltre, ci deve essere corrispondenza uno a uno tra gli elementi di Θ e quelli di $\mathcal{F}$; il modello parametrico deve essere, cioè, **identificabile**. Ovviamente, condizione essenziale perché una qualunque tecnica inferenziale parametrica possa essere efficace è che la vera legge generatrice dei dati, diciamo p_0, sia dentro $\mathcal{F}$, sia cioè un elemnto della classe. In altri termini, il modello deve essere **correttamente specificato**.

Generalmente, i dati osservati sono presentabili nella forma

$$\underline{y} = \{y_1, y_2, \ldots, y_n\},$$

dove $\underline{y}$ indica l'intera osservazione e il generico elemento y_i, $i = 1, 2, \ldots, n$, indica l'osservazione relativa alla i-esima tra le n unità campionarie. Spesso, tale osservazione (y_i) è essa stessa costituita da un ampio insieme di valori (per esempio un vettore).
Sia $\underline{Y} = \{Y_1, Y_2, \ldots, Y_n\}$, la variabile casuale di cui è realizzaione l'osservazione $\underline{y}$. Scegliere un modello statistico parametrico per l'osservazione campionaria, significa scegliere una classe di leggi $p_{\underline{Y}}(\underline{y}; \theta)$ per la variabile $\underline{Y}$, adeguate alla natura stessa dell'osservazione e "consistenti" con il problema che si deve trattare. In definitiva scegliere un modello pametrico equivale a individuare una classe di possibili *descrittori* della realtà complessa, che ha dato origine all'osservazione (intera) campionaria. Indicheremo un modello statistico parametrico come la classe

$$\mathcal{F} = \{p_{\underline{Y}}(\underline{y}; \theta),\ \theta \in \Theta\},$$

dove la funzione $p_{\underline{Y}}(\underline{y}; \theta)$ è definita su $\mathcal{Y} \times \Theta$, con $\mathcal{Y}$ detto **spazio campionario**. Lo spazio campionario è quindi l'insieme di tutti i possibili "valori" che l'intera osservazione può assumere. La funzione $p_{\underline{Y}}(\underline{y}; \theta)$ è detta **funzione del modello**. Per un modello correttamente specificato (identificabile) risulta $p_0 \in \mathcal{F}$; vuol dire che esisterà un valore del parametro, diciamo θ^0, tale che $p_0 = p_{\underline{Y}}(\underline{y}; \theta^0)$, per $\underline{y} \in \mathcal{Y}$. Spesso ci riferiremo a θ^0 come al **vero valore del parametro**.

Data l'osservazione campionaria $\underline{y}$ e scelto un modello statistico parametrico (identificabile e correttamente specificato), chiamiamo **funzione di verosimiglianza** per θ la funzione del modello $p_{\underline{Y}}(\underline{y}; \theta)$, vista come funzione di θ, fissata l'osservazione. Dal punto di vista della notazione scriveremo

$$L(\theta) = p_{\underline{Y}}(\underline{y}; \theta),$$

per $\theta \in \Theta$. Generalmente, $p_{\underline{Y}}(\underline{y}; \theta)$, per θ fissato, è data in termini di funzione di densità congiunta o funzione di probabilità congiunta. Quindi, $L(\theta) : \Theta \to \Re^+$.

Esempio 2.1 Supponiamo di disporre di dati provenienti da un esperimento che consiste nel lanciare n volte un dado. Immaginiamo di sospettare che il dado sia truccato e di voler controllare la fondatezza di questo sospetto. Immaginiamo che i dati siano forniti nella forma $\underline{y} = \{y_1, y_2, \ldots, y_n\}$, dove la generica osservazione y_i vale 1 se l'esito dell'i-esimo lancio è un numero pari, 0 altrimenti. Se il dato fosse perfettamente equilibrato, la probabilità di un esito pari, diciamo θ, sarebbe 1/2. Il parametro di interesse (e ovviamente ignoto) in questo problema è proprio θ. D'altra parte, la natura della singola risposta y_i, porta a ritenere che la generica variabile Y_i sia di Bernoulli, con parametro (probabilità di "successo") θ. Se assumiamo l'indipendenza degli esiti dei vari lanci, fissando la legge per la variabile marginale Y_i stiamo di fatto scegliendo un modello statistico parametrico

per l'intera osservazione, con funzione del modello (la funzione di probabilità congiunta relativa alla n-upla $\{Y_1, Y_2, \ldots, Y_n\}$): $p_{\underline{Y}}(\underline{y};\theta) = p_{Y_1}(y_1;\theta)p_{Y_2}(y_2;\theta)\ldots p_{Y_n}(y_n;\theta)$. Quindi,

$$L(\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \;=\; \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i}$$

è la funzione di verosimiglianza per θ, alla luce dell'osservazione $\underline{y}$. In questo caso, lo spazio parametrico è l'intervallo $(0,1)$. Lo spazio campionario è dato da $\mathcal{Y} = \{0,1\}\times\{0,1\}\times\ldots\times\{0,1\}$, n volte.

Esempio 2.2 Con riferimento all'Esempio 2.1, supponiamo che l'informazione relativa al risultato sperimentale sia sintetizzata da una singola osservazione, y, che fornisce il numero di volte in cui l'esito del lancio del dato sia stato un numero pari, nelle n prove. A parità delle condizioni sopra fissate, è naturale ritenere che la variabile Y, di cui è realizzazione y, sia binomiale di indice n e parametro θ, cioè, $Y \sim Bi(n,\theta)$. In questo caso, dunque, lo spazio campionario risulta essere l'insieme $\mathcal{Y} = \{0,1,2\ldots,n\}$, e il modello parametrico implicitamente fissato per l'intera (in questo caso unica) ossevazione y ha funzione del modello

$$p_Y(y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

Quindi la funzione di verosimiglianza per θ, relativa all'osservazione y, è $L_*(\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$, con y fissato, per $\theta \in (0,1)$.

Se teniamo conto (facendo riferimento alla notazione usata nell'Esempio 1) che $y = \sum_{i=1}^n y_i$, possiamo osservare che per la verosimiglianza deducibile da y vale la relazione

$$L_*(\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = \binom{n}{y}\theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} = \binom{n}{y}L(\theta).$$

Le due funzioni $L_*(\cdot)$ e $L(\cdot)$, intese come sole funzioni del parametro θ (date le osservazioni), sono tra loro proporzionali.

Esempio 2.3 Consideriamo i dati dell'esempio introduttivo, relativo ai voli della navicella Challenger. Nelle ipotesi formulate circa le coppie di osservazioni (y_i, t_i), per $i = 1,2,\ldots,23$, le variabili Y_i, di cui sono realizzazioni i valori y_i, sono (ciascuna condizionatamente al valore t_i) binomiali elementari con parametro $\pi_i = \frac{e^{\alpha+\beta t_i}}{1+e^{\alpha+\beta t_i}}$, indipendenti. Tutti i π_i dipendono, oltre che dai valori t_i della temperatura (che sono dati come noti), dalla coppia $\theta = (\alpha,\beta)$ di elementi ignoti. Se indichiamo con $\underline{Y} = \{Y_1, Y_2, \ldots, Y_{23}\}$ la variabile relativa allle osservazioni $y_1, y_2, \ldots, y_{23}$, le ipotesi formulate determinano di fatto la scelta di un modello parametrico $\mathcal{F}$ per l'osservazione che ha funzione del modello ottenuta come prodotto (data l'indipendenza) di contributi relativi alle singole coppie (y_i, t_i)

$$p_{\underline{Y}}(\underline{y};\theta) = p_{Y_1}(y_1;\theta)p_{Y_2}(y_2;\theta)\ldots p_{Y_{23}}(y_{23};\theta) = \prod_{i=1}^{23}[\pi_i(t_i;\theta)]^{y_i}[1-\pi_i(t_i;\theta)]^{1-y_i},$$

con $\pi_i(t_i;\theta) = e^{\alpha+\beta t_i}/[1+e^{\alpha+\beta t_i}]$. La funzione di verosimiglianza per θ, alla luce dell'osservazione campionaria $\underline{y}$, sarà dunque

$$L(\theta) = \prod_{i=1}^{23} [\pi_i(t_i;\theta)]^{y_i} [1-\pi_i(t_i;\theta)]^{1-y_i}.$$

Osservazione Nell'esempio 2.1, le variabili coinvolte $\{Y_1, Y_2, \ldots, Y_n\}$ sono indipendenti e di identica distribuzione (i.i.d.). L'osservazione generata, $\{y_1, y_2, \ldots, y_n\}$, costituisce un campione casuale semplice (c.c.s.). Nell'esempio 2.3, le variabili coinvolte sono ancora indipendenti ma non identicamente distribuite.

Esempio 2.4 Supponiamo di osservare, per n settimane, il numero di clienti che entrano in una filiale bancaria nell'orario di apertura del mercoledì. Siano $\{y_1, y_2, \ldots, y_n\}$ le osservazioni rilevate. Supponiamo di essere interessati ad aquisire informazioni sul numero medio di clienti che frequenta quella filiale il mercoledì. Se assumiamo che la generica osservazione possa essere pensata come realizzazione di una variabile Y_i di Poisson di parametro θ, e che le varabili siano tra loro indipendenti, l'osservazione campionaria costituisce un c.c.s., e per essa stiamo fissando un modello parametrico $\mathcal{F}$ con funzione del modello

$$p_{\underline{Y}}(\underline{y};\theta) = p_{Y_1}(y_1;\theta) p_{Y_2}(y_2;\theta) \ldots p_{Y_n}(y_n;\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!}.$$

Lo spazio campionario sarà in questo caso $\mathcal{Y} = \{0,1,2,3,\ldots\} \times \{0,1,2,3,\ldots\} \times \ldots \times \{0,1,2,3,\ldots\}$, n volte, e lo spazio prametrico sarà $\Theta = (0, +\infty)$. La funzione di verosimiglianza per θ, che nel problema posto rappresenta proprio l'entità di diretto interesse (**perché?**), risulta dunque essere

$$L(\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} = \theta^{\sum_{i=1} y_i} e^{-n\theta} \prod_{i=1}^{n} \frac{1}{y_i!},$$

per $\theta \in \Theta$.

0.3 Verosimiglianze equivalenti e statistiche sufficienti

Gli esempi 1 e 2 della sezione precedente si riferiscono allo stesso problema. Nel primo caso si suppone di avere l'informazione completa, cioè di disporre di tutti i dati relativi all'esperimento, $\{y_1, y_2, \ldots, y_n\}$; nel secondo caso ci si basa su una opportuna "sintesi", costituita dalla somma dei valori osservati $\sum_{i=1}^{n} y_i$. Il parametro che indicizza i due modelli coinvolti è lo stesso, θ, ed esso rappresenta anche l'entità di diretto interesse (probabilità di un risultato sul lancio che sia un numero pari). Le due funzioni di verosimiglianza, associate ai due modelli (che sono stati scelti in maniera **coerente**), hanno la caratteristica di essere tra loro proporzionali, quando considerate come funzioni di θ, fissatte le osservazioni. Due verosimiglianze che presentano questa caratteristica si dicono **equivalenti** tra loro, e, dal punto di vista inferenziale, portano alle stesse conclusioni sul parametro ignoto θ (e su qualunque altro elemento ignoto della poplazione di riferimento, che sarà necessariamente funzione di θ).
Ciò significa, quindi, che nel nostro caso non è necessario disporre di tutti i risultati dei singoli lanci

per fare infererenza su θ usando la funzione di verosimiglianza; è sufficiente conoscere solo il numero di volte in cui, su n lanci, il risultato è stato un numero pari. Una statistica (cioè, una funzione delle osservazioni $\{y_1, y_2, \ldots, y_n\}$), per la quale si sceglie un modello parametrico coerente con quello fissato per l'intera osservazione campionaria, che produce una verosimiglianza equivalente, si dice **statistica sufficiente** per l'inferenza sul parametro. Pertanto, con riferimento agli esempi 2.1 e 2.2, la statistica $\sum_{i=1}^n y_i$ (o equivalentemente l'oggetto casuale $\sum_{i=1}^n Y_i$ corrispondente) è statistica sufficiente per l'inferenza su θ. Tipicamente, una statistica sufficiente opera una riduzione dellla dimensione del problema, nel senso che il ricorso alla statistica sufficiente produce una riduzione della dimensione dello spazio campionario. Nei due esempi, la dimensione di $\mathcal{Y}$ passa, in effetti, da n a 1 (la statistica sufficiente è unidimensionale).

L'esempio 4 nella sezione precedente, ci indica un modo per individuare in pratica una statistica sufficiente. Infatti, in quell'esempio, per la funzione di verosimiglianza per il parametro θ, che caratterizza la legge di Poisson, vale la relazione

$$L(\theta) = \prod_{i=1}^n \frac{\theta^{y_i} e^{-\theta}}{y_i!} = \theta^{\sum_{i=1} y_i} e^{-n\theta} \prod_{i=1}^n \frac{1}{y_i!} \;\propto\; \theta^{\sum_{i=1} y_i} e^{-n\theta},$$

dove, al solito, la proporzionalità è intesa rispetto a funzioni di θ. Se poniamo $L_*(\theta) = \theta^{\sum_{i=1}} e^{-n\theta}$, $L(\theta)$ e $L_*(\theta)$ sono verosimiglianze equivalenti per l'inferenza su θ. Inoltre, $L_*(\theta)$ non è ulteriormente "riducibile" (escludendo ulteriori "fattori di proporzionalità") e dipende dai dati osservati $\{y_1, y_2, \ldots, y_n\}$ solo tramite la statistica $\sum_{i=1}^n y_i$. In altri termini, per scrivere la funzione di verosimiglianza per θ non occorre conoscere i valori di tutte le osservazioni ma basta il valore della loro somma: $\sum_{i=1}^n y_i$ è dunque statistica sufficiente.
In generale, data l'osservazione campionaria $\underline{y}$ e scelto un modello parametrico $\mathcal{F}$ (identificabile e correttamente specificato), indicizzato da un parametro θ (diciamo), una volta scritta la funzione di verosimiglianza per θ nella sua forma equivalente non ulteriormente riducibile (che potremmo chiamare **nucleo della verosimiglianza**), se essa dipende dai dati solo tramite una qualche statistica (funzione) di sintesi, tale statistica è statistica sufficiente per l'inferenza su θ. [1]
Dovrebbe essere chiaro, a questo punto, che per un modello parametrico, scelto con riferimento ad una certa osservazione campionaria, non esiste una sola funzione di verosimiglianza; piuttosto esiste una classe costituita da una infinità di funzioni tutte equivalenti tra loro: in pratica si userà la forma più comoda, spesso quella che rappresenta l'elemento che abbiamo chiamato nucleo. Infine, ogni funzione biunivoca di una statistica sufficiente è essa stessa statistica sufficiente (giusto per illustrare, nell'esempio 2.4, $e^{\sum_{i=1}^n y_i}$ e $3 + 4\sum_{i=1}^n y_i$ sono ancora statistiche sufficienti).

0.4 Parametrizzazioni

In un contesto parametrico, ogni operazione di inferenza su una qualunque caratteristica del meccanismo casuale che genera i dati o, in altri termini, su una qualunque entità di interesse relativa al fenomeno che si sta studiando nella popolazione di riferimento, passa attraverso l'inferenza sul parametro che indicizza il modello (identificabile e correttamente specificato) scelto per rappresentare la realtà. A ben vedere, il parametro rappresenta una sorta di "etichetta", che distingue i vari

[1] Questo modo di procedere equivale al cosiddetto criterio di fattorizzazione di Neyman-Fisher.

elementi della classe parametrica $\mathcal{F}$. A volte, per raggiungere un determinato obiettivo inferenziale, può essere conveniente usare una "etichettatura" piuttosto che un'altra.

Riprendiamo l'esempio 4 della sezione 2. Supponiamo di essere interessati non tanto a valutare il numero medio di clienti che frequenta la filiale bancaria il mercoledì, ma piuttosto la probabilità che al mercoledi si presenti in filiale almeno un cliente. Dato il modello scelto, tale probabilità, ignota, è funzione di θ (parametro della Poisson) e vale $1 - \Pr\{Y_i = 0\} = 1 - e^{-\theta}$. Posto $\psi = 1 - e^{-\theta}$, per $\theta > 0$ ψ è funzione monotona (strettamente) crescente (**perché?**) di θ. Quindi $\psi = g(\theta)$ è funzione biunivova, cioè uno-a-uno, di θ. Questo implica che l'elemento della classe parametrica $\mathcal{F}$, corrispondente alla etichetta θ_1, per esempio, potremmo indicarlo usando una nuova etichetta che è $\psi_1 = g(\theta_1)$. Se effettuiamo questo cambio di etichetta per tutti gli elementi di $\mathcal{F}$, diciamo ch stiamo facendo un'operazione di **riparametrizzazione**.
In generale, **cambiare parametrizzazione** significa passare da un modello di partenza, indicizzato da un parametro, diciamo θ, ad un "nuovo" modello indicizzato da un nuovo parametro, diciamo ψ, che è legato a θ da una relazione biunivoca $\psi = g(\theta)$. Se $p_{\underline{Y}}(\underline{y}; \theta)$ è la funzione del modello di partenza, la funzione del modello nella nuova parametrizzazione sarà semplicemente data da

$$p^{\psi}_{\underline{Y}}(\underline{y}; \psi) = p_{\underline{Y}}(\underline{y}; g^{-1}(\psi)),$$

dove g^{-1} rappresenta la funzione inversa di g, e il suffisso ψ a esponente è un elemento di notazione, che serve a distinguere la funzione del modello nella nuova parametrizzazione dalla vecchia. Ancora a livello di notazione, il modello

$$\mathcal{F} = \{p_{\underline{Y}}(\underline{y}; \theta),\ \theta \in \Theta\}$$

nella parametrizzazione di partenza, diventa

$$\mathcal{F}^{\psi} = \{p^{\psi}_{\underline{Y}}(\underline{y}; \psi),\ \psi \in \Psi\},$$

nella nuova. Il nuovo spazio parametrico Ψ è ovviamente legato a Θ, in quanto immagine di Θ attraverso la funzione $g(\cdot)$: $\Psi = \{\psi :\ \psi = g(\theta),\ \ \theta \in \Theta\}$.
Nel caso dell'esempio 4 della sezione 2, si ha $\Psi = (0, 1),\ \ g^{-1}(\psi) = -\log(1 - \psi)$ e

$$p^{\psi}_{\underline{Y}}(\underline{y}; \psi) = \prod_{i=1}^{n} \frac{[-\log(1-\psi)]^{y_i} e^{\log(1-\psi)}}{y_i!}.$$

Si osservi che, con questa parametrizzazione, il parametro che indicizza il modello diventa l'entità di diretto interesse. Per la funzione di verosimiglianza per il nuovo parametro ψ vale la relazione

$$L^{\psi}(\psi) = p^{\psi}_{\underline{Y}}(\underline{y}; \psi) = p_{\underline{Y}}(\underline{y}; g^{-1}(\psi)) = L(g^{-1}(\psi)),$$

dove $L(\cdot)$ indica la funzione di verosimiglianza per θ. In sostanza, la verosimiglianza per il nuovo parametro ψ si può ottenere dalla verosimiglianza per il parametro di partenza θ esplicitando, nella espressione della seconda, la relazione tra θ e ψ.

0.5 Stima e stimatore di massima verosimiglianza

Finora abbiamo sempre considerato la funzione di verosimiglianza per un parametro θ, che indicizza un modello statistico parametrico $\mathcal{F}$, scelto con riferimento all'osservazione $\underline{y}$, come funzione del

parametro data l'osservazione. In realà abbiamo detto che la funzione di verosimiglianza coincide con la funzione del modello. Quindi, più correttamente, dovremmo scrivere

$$L(\theta; \underline{y}) = p_{\underline{Y}}(\underline{y}; \theta).$$

Questa espressione ci porta a fare una riflessione sulla natura della funzione di verosimiglianza. In effetti, *a priori* rispetto all'esperimento che genera i dati (l'osservazione campionaria), $L(\theta; \underline{y})$ ci dà indicazioni sulla "probabilità" con cui un elemento della classe $\mathcal{F}$ (si pensi a θ fissato) può generare un qualunque possibile dato $\underline{y}$; nel caso di variabili discrete, $p_{\underline{Y}}(\theta; \underline{y})$ è proprio tale probabilità. *A posteriori*, cioè una volta in possesso dell'osservazione $\underline{y}$, $L(\theta; \underline{y})$ ci fornisce indicazioni sulla "plausibilità" dei vari elementi della classe (si pensi a θ che varia) come elementi generatori del dato osservato. In questo senso, alla luce del dato osservato, la verosimiglianza rappresenta una misura della **plausibilità** che possiamo associare ai vari elementi della classe $\mathcal{F}$. Se accettiamo questa idea, viene quasi automatica l'individuazione di un criterio per risolvere il problema di stima puntuale usando come strumento la funzione di verosinmiglianza.

Abbiamo già detto che nell'approccio parametrico, una volta scelto il modello statistico $\mathcal{F}$, l'unico elemento ignoto nella formalizzazione è rappresentato dal parametro θ che indicizza il modello stesso. È chiaro quindi che qualsiasi operazione di inferenza, per qualunque entità di interesse, dovrà, in qualche modo, "passare" per l'inferenza su θ. Alla luce di quanto detto sopra, dovrebbe apparire del tutto ragionevole, almeno in linea di principio, il seguente criterio per ottenere una stima (puntuale) di θ: si sceglie come stima del parametro quel valore che massimizza la funzione di verosimiglianza, cioè che etichetta l'elemento della classe che più verosimilmente ha generato l'osservazione $\underline{y}$. Se indichiamo con $\hat{\theta}$ questo valore (assumendo che esista), esso sarà tale che

$$L(\hat{\theta}) \geq L(\theta) \quad \text{per ogni} \quad \theta \in \Theta.$$

Chiameremo $\hat{\theta}$ **stima di massima verosimiglianza** di θ. Dato che la verosimiglianza dipende dai dati, anche $\hat{\theta}$ dipenderà dai dati (in ogni problema ben posto, l'inferenza si basa sui dati osservati). Per sottolineare questo aspetto, quando occorre, possiamo scrivere $\hat{\theta}(\underline{y})$; la stima è, come deve essere, una statistica. Dato che l'osservazione $\underline{y}$ è realizzazione di un oggetto casuale $\underline{Y}$, anche la stima $\hat{\theta}$ sarà realizzazione di una variabile casuale che possiamo indicare con $\hat{\theta}(\underline{Y})$; a tale variabile ci riferiremo come allo **stimatore di massima verosimiglianza** per il parametro θ.

0.6 Stima di massima verosimiglianza: aspetti computazionali

Dal punto di vista pratico, per com'è definita, la stima di massima verosimiglianza può presentare problemi di esistenza e unicità. Ciò vuol dire che ci potrebbero essere situazioni in cui la stima non esiste (la funzione di verosimiglianza non ammette punto di massimo), oppure situazioni in cui esiste ma non è unica (la verosimiglianza ha più punti di massimo assoluto). In quest'ultimo caso, spetta allo statistico, se la cosa è ragionevole, procedere a una opportuna scelta della stima tra i candidati possibili, accettando, di fatto, una complicazione del problema e l'idea che eventuali successive procedure di inferenza dovranno avere adeguata giustificazione, anche dal punto di vista teorico.

Quando la verosimiglianza ha un unico punto di massimo assoluto, l'individuazione di $\hat{\theta}$ comporta al più problemi di natura computazionale. Come mostrato dagli esempi che seguono, l'individuazione del punto di massimo di $L(\theta)$ può richiedere solo un'ispezione dei valori che $L(\cdot)$ può assumere (questo in casi molto semplici, in cui lo spazio parametrico Θ è costituito da un numero finito di punti), oppure il ricorso a metodi dell'Analisi Matematica, di tipo diverso a seconda della natura stessa della funzione. In molte situazioni, piuttosto che lavorare direttamente sulla verosimiglianza, si preferisce, per comodità, lavorare su una trasformata monotona che è $l(\theta) = \log(L(\theta))$. Tale funzione è chiamata **funzione di log-verosimiglianza**. Se $\hat{\theta}$ è punto di massimo di $L(\theta)$, lo sarà anche di $l(\theta)$.

Esempio 6.1 Si pensi ad un esperimento bernoulliano, consistente in $n = 5$ prove indipendenti, con probabilità di successo, θ, ad ogni prova. Si supponga che θ possa assumere solo i valori 0,2, 0,4, 0,6 e 0,8. Vogliamo ottenere una stima di θ, sulla base del risultato sperimentale y che fornisce il numero di successi osservati.
La natura della variabile Y, di cui y è realizzazione, è chiara: si tratta di una variabile binomiale, di indice 5 e parametro θ, cioè $Y \sim Bi(5, \theta)$. Lo spazio campionario $\mathcal{Y}$ è l'insieme $\{0, 1, 2, 3, 4, 5\}$, mentre, sulla base delle conoscenze preliminari, lo spazio parametrico è $\Theta = \{0,2, 0,4, 0,6, 0,8\}$.

y	$\theta = 0,2$	$\theta = 0,4$	$\theta = 0,6$	$\theta = 0,8$
0	0,32768	0,07776	0,01024	0,00032
1	0,4096	0,2592	0,0768	0,0064
2	0,2048	0,3456	0,2304	0,0512
3	0,0512	0,2304	0,3456	0,2048
4	0,0064	0,0768	0,2592	0,4096
5	0,00032	0,01024	0,07776	0,32768

La tabella a fianco riporta i valori della funzione del modello $p_Y(y;\theta) = \binom{5}{y}\theta^y(q1-\theta)^{5-y}$, al variare di y in $\mathcal{Y}$ e di θ in Θ. Essa riporta, quindi, i valori della verosimiglianza. La stima di massima verosimiglianza può essere individuata banalmente per ispezione: se il risultato sperimentale fosse $y = 2$, la funzione di verosimiglianza per $\theta \in \Theta$ sarebbe fornita dalla terza riga della tabella, e il valore massimo si avrebbe in corrispondenza del valore 0,4 per θ. Pertanto, avremmo $\hat{\theta} = 0,4$.

Esempio 6.2 In un'officina, una macchina deve effettuare dei fori su lastre metalliche, di diametro nominale 15 centimetri. Come ogni macchina, anche quella in oggetto può lavorare con un errore, che risulta accettabile fintanto che si mantiene contenuto. Si supponga di non avere nessuna informazione *a priori* sull'errore che la macchina commette. L'utente è interessato ad avere una valutazione statistica dell'errore assoluto massimo con cui essa può operare. Per ottenere tale valutazione si procede a raccogliere, a intervalli casuali, n lastre forate, e su ciascuna lastra si misura l'errore assoluto sul valore del diametro del foro, rispetto al valore nominale. L'osservazione campionaria è costituita dalla n-upla $\underline{y} = \{y_1, y_2, \ldots, y_n\}$, che pensiamo realizzazione di una variabile $\underline{Y} = \{Y_1, Y_2, \ldots, Y_n\}$. Data l'assenza di informazioni preliminari e visto lo scopo dell'esperimento, possiamo assumere che la generica osservazione y_i sia realizzazione di una variabile Y_i uniforme su $(0, \theta]$, cioè $Y_i \sim U(0, \theta]$. Se assumiamo che le variabili Y_i siano indipendenti tra loro, il campione è un c.c.s. e, di fatto, stiamo fissando un modello parametrico $\mathcal{F}$ per l'intera osservazione con funzione del modello (**perché?**)

$$p_{\underline{Y}}(\underline{y};\theta) = \prod_{i=1}^{n} \frac{1}{\theta} I_{(0,\theta]}(y_i),$$

dove $I_A(x)$ è la **funzione indicatrice** dell'evento $x \in A$, con $\theta \in \Theta = (0, +\infty)$. Data l'osservazione $\underline{y}$, la funzione di verosimiglianza per θ è

$$L(\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} I_{(0,\theta]}(y_i) = \frac{1}{\theta^n} \prod_{i=1}^{n} I_{[y_i,+\infty)}(\theta).$$

Il prodotto di funzioni indicatrici è ancora una funzione indicatrice, che vale uno se e solo se valgono 1 tutte le indicatrici coinvolte. Nel nostro caso, tutte le osservazioni y_i devono essere minori o uguali di θ, e questo equivale a richiedere che la più grande tra le osservazioni, $y_{(n)}$, soddisfi lo stesso vincolo. In definitiva, risulta che

$$L(\theta) = \frac{1}{\theta^n} I_{[y_{(n)},+\infty)}(\theta) = \left\{ \begin{array}{ll} \frac{1}{\theta^n} & \text{se } \theta \geq y_{(n)} \\ 0 & \text{altrove} \end{array} \right\}.$$

Quindi, la funzione di verosimiglianza non è continua ed è strettamente decrescente su $[y_{(n)}, +\infty)$, cioè laddove non è nulla. Ne deriva che la stima di massima verosimiglianza di θ è $\hat{\theta} = y_{(n)}$. La variabile casuale corrispondente, diciamo $Y_{(n)}$, è lo stimatore di massima verosimiglianza. Si osservi che, in questo caso, $Y_{(n)}$ è anche statistica sufficiente per l'inferenza su θ.

Esempio 6.3 Riconsideriamo il problema trattato nell'esempio 2.4. Se l'interesse è qello di acquisire informazioni sul numero medio di clienti che frequenta la filiale al mercoledì, possiamo risolvere recuperando una stima del parametro θ, che nella formalizzazione considerata rappresenta proprio l'entità di interesse. Per la funzione di verosimiglianza vale la relazione

$$L(\theta) = \theta^{\sum_{i=1}^n y_i} e^{-n\theta} \prod_{i=1}^{n} \frac{1}{y_i!} \propto \theta^{\sum_{i=1}^n y_i} e^{-n\theta}.$$

Se passiamo al logaritmo, otteniamo, per la funzione di log-verosimiglianza,

$$l(\theta) = \log(\theta) \sum_{i=1}^{n} y_i - n\theta + c,$$

dove c rappresenta una costante (rispetto all'argomento θ) additiva. Da ciò possiamo intanto dedurre un risultato generale: **verosimiglianze equivalenti**, per come le abbiamo definite in precedenza, **generano log-verosimiglianze equivalenti, nel senso di funzioni di θ che differiscono tra loro per costanti (rispetto a θ) additive**. Inoltre se scegliamo come log-verosimiglianza la formulazione meno complessa $l(\theta) = \log(\theta) \sum_{i=1}^n y_i - n\theta$, possiamo osservare che tale funzione, definita su $\Theta = (0, +\infty)$, è continua e derivabile più volte. In altri termini, è una funzione sufficientemente regolare di θ, quindi il suo punto di massimo può essere cercato tra i suoi punti di stazionarietà. Derivando rispetto a θ, si ottiene

$$l_*(\theta) = \frac{d\ l(\theta)}{d\theta} = \frac{\sum_{i=1}^n y_i}{\theta} - n$$

da cui, uguagliando a zero e risolvendo in θ, si ottiene l'unico punto di stazionarietà quando $\theta = \frac{1}{n} \sum_{i=1}^n y_i$. Dato che la derivata seconda della log-verosimiglianza vale $-\frac{\sum_{i=1}^n y_i}{\theta^2}$ che è sempre

negativa su tutto Θ (la log-verosimiglianza è funzione strettamente concava), il punto di stazionarietà trovato è anche l'unico punto di massimo. Quindi la stima di massima verosimiglianza di θ è la media campionaria: $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i$. In questo caso, l'approccio basato sulla verosimiglianza fornisce lo stesso stimatore di θ che si otterrebbe usando il metodo dei momenti **(perché?)**.

Esempio 6.4 In un sito archeologico, in una regione abitata in un passato remoto dai membri di una tribù estinta da secoli, viene rinvenuta un'area cimiteriale. I ricercatori raccolgono un campione di n femori di soggetti adulti e ne misurano la lunghezza. Gli studiosi sono interessati ad ottenere una valutazione statistica (una stima) della lunghezza media del femore dei soggetti adulti di quell'antica popolazione.
Se assumiamo (come spesso succede con le misure antropometriche) che la generica misura y_i sia realizzazione di una variabile casuale normale, e che le misure siano indipendenti, l'osservazione campionaria $\underline{y} = \{y_1, y_2, \ldots, y_n\}$, che pensiamo realizzazione di una variabile $\underline{Y} = \{Y_1, Y_2, \ldots, Y_n\}$, costituisce un c.c.s., con $Y_i \sim N(\mu, \sigma^2)$. Così facendo, stiamo associando all'intera osservazione $\underline{y}$ un modello statistico parametrico $\mathcal{F}$ con funzione del modello

$$p_{\underline{Y}}(\underline{y};\theta) = p_{Y_1}(y_1;\theta)p_{Y_2}(y_2;\theta)\ldots p_{Y_n}(y_n;\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2},$$

con $\theta = (\mu, \sigma^2)$ parametro bidimensionale che varia nello spazio parametrico $\Theta = \Re \times (0, +\infty)$ **(qual è lo spazio campionario?)**. La funzione di verosimiglianza per θ, data l'osservazione $\underline{y}$, è

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} \propto (1/\sigma^2)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}.$$

Passando alla log-verosimiglianza, si ottiene

$$l(\theta) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2.$$

Questa espressione ci permette di calcolare la stima di massima verosimiglianza di θ con un certa semplicità. Infatti se in essa fissiamo σ^2, vediamo che per massimizzarla in μ bisogna minimizzare la funzione $\sum_{i=1}^{n}(y_i-\mu)^2$. Da una nota proprietà della media campionaria (minimizza la somma dei quadrati degli scarti), deduciamo che il valore di μ che massimizza $l(\theta)$ è $\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, a prescindere dal valore fissato di σ^2. A questo punto, per ottenere la stima di σ^2, basta massimizzare la funzione di una sola variabile

$$\ell(\sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

Si tratta di una funzione regolare, cioè continua e derivabile più volte, e il suo punto di massimo può essere cercato tra i punti di stazionarietà. Derivando rispetto a σ^2, otteniamo

$$\frac{d\,\ell(\sigma^2)}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

Uguagliando a zero e risolvendo l'equazione in σ^2, si ricava un punto di stazionarietà di $\ell(\sigma^2)$, che corrisponde a $\sigma^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2$. Derivando ancora, si ottiene che la derivata seconda è pari a

$$\frac{n}{2\sigma^4} - \frac{n\hat{\sigma}^2}{\sigma^6} = \frac{n}{2\sigma^4}\left(1 - \frac{2\hat{\sigma}^2}{\sigma^2}\right),$$

dove si è posto $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2$, e che, come si può notare, è negativa quando $\sigma^2 = \hat{\sigma}^2$. Quindi, la stima di massima verosimiglianza di σ^2 è proprio la varianza campionaria $\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2$. In definitiva, concludiamo che $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = \left(\bar{y}, \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2\right)$. La variabile casuale (bidimensionale) corrispondente $\left(\bar{Y}, \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2\right)$, definisce lo stimatore di massima verosimiglianza. In questo caso, lo stimatore di massima verosimiglianza per media e varianza del carattere considerato nella popolazione di riferimento, ripropone ciò che avremmo ottenuto ragionando per analogia. Si osservi infine che, in questo particolare problema, è μ a rappresentare il parametro di interesse. L'altro elemento di θ, σ^2, è comunque supposto ignoto nella nostra formalizzazione ma non è di diretto interesse. Si dice, in casi come questo, che un tale parametro è un **parametro di disturbo**.

Negli esempi sopra considerati, si riesce sempre ad ottenere stima e stimatore di massima verosimiglianza in forma analitica. Purtroppo non è sempre così e spesso, specie nei casi più complessi, bisogna procedere per via numerica, massimizzando direttamentela verosimiglianza $L(\theta)$ utilizzando qualche algoritmo di ottimizzazione numerica, o risolvendo, nei casi di verosimiglianza regolare, l'equazione $l_*(\theta) = 0$ con algoritmi numerici. Tale equazione è detta **equazione (o sistema di equazioni) di verosimiglianza**. Diremo che un modello statistico parametrico $\mathcal{F}$ **ha verosimiglianza regolare se lo spazio parametrico Θ è un sottoinsieme aperto dello spazio Euclideo di una qualche dimensione (finita) $k \geq 1$, e se la log-verosimiglinza associata, $l(\theta)$, è derivabile almeno tre volte, con derivate continue in Θ.**

Esempio 6.5 In un esperimento sull'efficacia di una sostanza insetticida per zanzare, n insetti sono collocati in apposite teche (una per ogni insetto) di uguale dimensione. Nelle teche viene immessa una certa quantità di veleno e, per ogni zanzara si registra la durata di vita (in minuti) fino alla morte. Sia $\underline{y} = \{y_1, y_2, \ldots, y_n\}$ l'osservazione campionaria, sulla base della quale si vuole valutare la durata media di vita degli insetti avvelenati. Per procedere si assuma che la generica variabile coinvolta Y_i abbia funzione di densià $p_{Y_i} = (y_i; \theta) = \theta y_i^{\theta-1} e^{-y_i^\theta}$, per $y_i > 0$ e $\theta > 0$, e che ci sia indipendenza tra le morti degli insetti. All'intera osservazione $\underline{y}$ rimane allora associato un modello statistico parametrico $\mathcal{F}$, con funzione del modello

$$p_{\underline{y}}(\theta; \underline{y}) = \prod_{i=1}^n \theta y_i^{\theta-1} e^{-y_i^\theta},$$

per $\theta \in \Theta = (0, +\infty)$ **(qual è lo spazio campionario?)**. In questa formalizzazione, l'entità di diretto interesse è il valore atteso di Y_i, cioè la durata media di vita di una zanzara avvelenata. Si può verificare che tale quantità è funzione del parametro θ che indicizza il modello. Più precisamente, risulta $E(Y_i) = \frac{\Gamma(1/\theta)}{\theta}$, in cui $\Gamma(\cdot)$ indica la funzione gamma.[2] Come ottenere la stima di massima

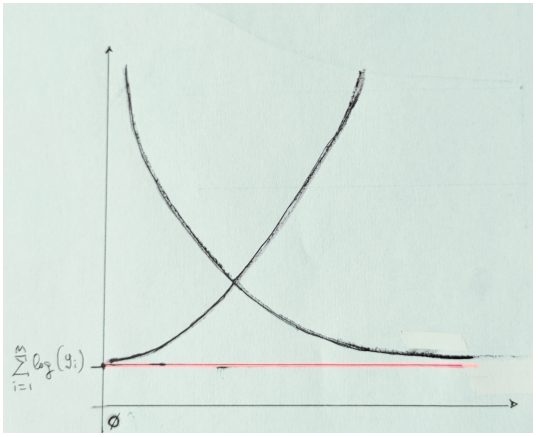
[2] $$E(Y_i) = \int_0^{+\infty} \theta y^\theta e^{-y^\theta} dy = \int_0^{+\infty} \theta t e^{-t} \frac{1}{\theta} t^{\frac{1}{\theta}-1} dt = \int_0^{+\infty} t^{\frac{1}{\theta}} e^{-t} dt = \int_0^{+\infty} \frac{\Gamma(\frac{1}{\theta}+1)}{\Gamma(\frac{1}{\theta}+1)} t^{\frac{1}{\theta}} e^{-t} dt,$$

verosimiglianza di θ ? La funzione di verosimiglianza è regolare, quindi possiamo considerare la funzione di log-verosimiglianza e cercare il suo punto di massimo tra i punti di stazionarietà. Risulta

$$l(\theta) = n\log(\theta) + (\theta - 1)\sum_{i=1}^{n}\log(y_i) - \sum_{i=1}^{n} y_i^{\theta},$$

la cui derivata, rispetto a θ, è

$$l_*(\theta) = \frac{n}{\theta} + \sum_{i=1}^{n}\log(y_i) - \sum_{i=1}^{n} y_i^{\theta}\log(y_i).$$



Risolvere l'equazione di verosimiglianza $l_*(\theta) = 0$, significa trovare i valori di θ per cui le due funzioni $\frac{n}{\theta} + \sum_{i=1}^{n}\log(y_i)$ e $\sum_{i=1}^{n} y_i^{\theta}\log(y_i)$ si uguagliano. Come funzioni di θ, tali funzioni sono l'una strettamente decrescente, l'altra strettamente crescente. Quindi, come mostra il grafico a fianco, esiste un un'unica radice dell'equazione, ovvero un unico punto di stazionarietà. Inoltre, la derivata seconda della log-verosimiglianza vale

$$-\frac{n}{\theta^2} - \sum_{i=1}^{n} y_i^{\theta}[\log(y_i)]^2,$$

ed è sempre (strettamente) negativa, per ogni $\theta \in \Theta$. Pertanto l'unico punto di stazionarietà è anche punto di massimo.

La radice dell'equazione di verosimiglianza non può però essere ottenuta analiticamente. È necessario utilizzare un opportuno algoritmo numerico, come, ad esempio, **l'algoritmo di Newton-Raphson**. Scelto un punto (iniziale) θ_0, localmente vale lo sviluppo di Taylor $l_*(\theta) = l_*(\theta_0) + (\theta - \theta_0)l_{**}(\theta_0) + resto$ che porta all'approssimazione $l_*(\theta) = l_*(\theta_0) + (\theta - \theta_0)l_{**}(\theta_0)$ in un intorno di θ_0. Qui $l_{**}(\cdot)$ indica la derivata seconda della log-verosimiglianza. Tale funzione approssimante è niente altro che la retta tangente la funzione $l_*(\theta)$ nel punto di coordinate $(\theta_0, l_*(\theta_0))$. Ponendo $l_*(\theta_0) + (\theta - \theta_0)l_{**}(\theta_0) = 0$ e risolvendo in θ, si ottiene il nuovo punto (che corrisponde all'intersezione della retta tangente con l'asse delle ascisse)

$$\theta_1 = \theta_0 - \frac{l_*(\theta_0)}{l_{**}(\theta_0)}.$$

Si riparte quindi dal punto θ_1 per ottenere quello successivo $\theta_2 = \theta_1 - \frac{l_*(\theta_1)}{l_{**}(\theta_1)}$, e così via fino a quando la differenza tra due soluzioni successive trovate, diciamo θ_s e θ_{s+1}, è sufficientemente piccola, secondo qualche criterio fissato. Se il punto iniziale θ_0 è scelto adeguatamente (tipicamente non deve

dopo aver operato la sostituzione $t = y^{\theta}$. Tenendo conto che $\frac{1}{\Gamma(\frac{1}{\theta}+1)} t^{\frac{1}{\theta}} e^{-t}$ è la densità di una variabile gamma di parametri $(\frac{1}{\theta} + 1)$ e 1, si ottiene $E(Y_i) = \Gamma(\frac{1}{\theta} + 1) = \frac{\Gamma(1/\theta)}{\theta}$.

essere "troppo" distante dalla radice cercata) l'algoritmo converergerà alla soluzione dell'equazione di verosimiglianza, e il valore della stima di massima verosimiglianza $\hat{\theta}$ sarà approssimativamente θ_s. Nel caso in esame,

$$l_*(\theta) = \frac{n}{\theta} + \sum_{i=1}^{n} \log(y_i) - \sum_{i=1}^{n} y_i^{\theta} \log(y_i) \quad \text{e} \quad l_{**}(\theta) = -\frac{n}{\theta^2} - \sum_{i=1}^{n} y_i^{\theta} [\log(y_i)]^2.$$

Come esercizio, si ponga $n = 4$ con $\underline{y} = \{1,18,\ 0,47,\ 1,09,\ 1,34\}$. Si tracci la funzione di log-verosimiglianza per θ e si calcolino le approssimazioni successive per $\hat{\theta}$, fornite dai primi 4 passi dell'algoritmo di Newton-Raphson (scegliendo il punto iniziale).

Esempio 6.4 (ripresa) Nell'esempio 6.4, la verosimiglianza per il parametro bidimensionale $\theta = (\mu, \sigma^2)$, relativa all'osservazione $\underline{y} = \{y_1, y_2, \ldots, y_n\}$, è

$$L(\theta) = \propto (1/\sigma^2)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2} = (1/\sigma^2)^{n/2} e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} y_i^2 - 2\mu\sum_{i=1}^{n} y_i + n\mu^2\right]}.$$

Questa espressione non è ulteriormente riducibile eliminando fattori di proporzionalità, e mostra che la coppia $(\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} Y_i^2)$ è statistica sufficiente per l'inferenza su θ. Naturalmente, anche l_* è in questo caso un vettore di dimensione due:

$$l_*(\theta) = \begin{pmatrix} \frac{\partial l(\theta)}{\partial \mu} \\ \frac{\partial l(\theta)}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} -\frac{n\mu}{\sigma^2} + \frac{\sum_{i=1}^{n} y_i}{\sigma^2} \\ -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^4} \end{pmatrix}.$$

Abbiamo già individuato la stima di massima verosimiglianza $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$, con $\hat{\mu} = \bar{y}$ e $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$, e sappiamo che $l_*(\hat{\theta}) = 0$. Se il parametro è bidimensionale come in questo caso, avreo una matrice di derivate seconde di dimensione 2×2. In particolare,

$$l_{**}(\theta) = \begin{pmatrix} \frac{\partial^2 l(\theta)}{\partial \mu^2} & \frac{\partial^2 l(\theta)}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 l(\theta)}{\partial \mu \partial \sigma^2} & \frac{\partial^2 l(\theta)}{\partial (\sigma^2)^2} \end{pmatrix} = \begin{pmatrix} -\frac{n}{\sigma^2} & \frac{n\mu}{\sigma^4} - \frac{\sum_{i=1}^{n} y_i}{\sigma^4} \\ \frac{n\mu}{\sigma^4} - \frac{\sum_{i=1}^{n} y_i}{\sigma^4} & \frac{n}{2\sigma^4} - \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{\sigma^6} \end{pmatrix},$$

ed è facile verificare che

$$l_{**}(\hat{\theta}) = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2\hat{\sigma}^4} \end{pmatrix}.$$

Quindi, nel punto $\hat{\theta}$ il vettore delle derivate della log-verosimiglianza è nullo e la matrice delle derivate seconde è caratterizzata dal fatto che l'elemento nella posizione 1,1 è negativo e il suo determinante è positivo: formalmente, $[l_{**}(\hat{\theta})]_{11} < 0$ e $\det[l_{**}(\hat{\theta})] > 0$. **Queste tre condizioni caratterizzano ogni punto di massimo (almeno locale), nel caso di parametro di dimensione due e verosimiglianza regolare**.

Esercizio. Con riferimento all'Esempio 2.3, si ottenga la stima di massima verosimiglianza del parametro $\theta = (\alpha, \beta)$ nel modello scelto per i dati relativi ai voli dello Shuttle.

Esempio 6.6 Si supponga di disporre di una osservazione campionaria, costituita dalla n-upla $\underline{y} = \{y_1, y_2, \ldots, y_n\}$, che pensiamo realizzazione di una variabile $\underline{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ a componenti

Y_i indipendenti e di identica distribuzione, uniforme su $[\theta - a, \theta + a]$. Qui $a > 0$ è un valore fissato (noto) e $\theta \in (-\infty, +\infty)$ è il parametro ignoto. La funzione del modello parametrico risultante da queste assunzioni è

$$p_{\underline{Y}}(\underline{y}; \theta) = \prod_{i=1}^{n} \frac{1}{2a} I_{[\theta-a, \theta+a]}(y_i),$$

dove $I_A(x)$ è la funzione indicatrice dell'evento $x \in A$. Data l'osservazione $\underline{y}$, la funzione di verosimiglianza per θ è

$$L(\theta) = \frac{1}{(2a)^n} \prod_{i=1}^{n} I_{[\theta-a, \theta+a]}(y_i),$$

per $\theta \in \Theta = (-\infty, +\infty)$. La verosimiglianza è non nulla solo se tutte le funzioni indicatrici valgono 1, e può essere quindi riscritta nella forma seguente:

$$L(\theta) = \frac{1}{(2a)^n} I_{(-\infty, y_{(1)}+a]}(\theta) I_{[y_{(n)}-a, +\infty]}(\theta) = \left\{ \begin{array}{cc} \frac{1}{(2a)^n} & \text{se } \; y_{(n)} - a \leq \theta \leq y_{(1)} + a \\ 0 & \text{altrove} \end{array} \right\},$$

dove $y_{(1)}$ e $y_{(n)}$ indicano, rispettivamente, il più piccolo e il più grande tra i valori osservati. Pertanto, la funzione di verosimiglianza non solo non è continua ma è costante laddove non è nulla. Ne deriva che la stima di massima verosimiglianza di θ non esiste unica. Può essere ragionevole, in questo caso, adottare (scegliere) come stima di massima verosimiglianza la semisomma dei due estremi dell'intervallo in cui la verosimiglianza assume il valore comune $\frac{1}{(2a)^n}$. Ne risulta $\hat{\theta} = (y_{(1)} + y_{(n)})/2$.[3] Lo stimatore di massima verosimiglianza corrispondente sarà rappresentato dalla variabile casuale $(Y_{(1)} + Y_{(n)})/2$.

Esempio 6.7 Sia ora $\underline{y} = \{y_1, y_2, \ldots, y_n\}$, un c.c.s. di dimensione n da una variabile continua con funzione di densità $p_Y(y; \theta) = (2/\pi)\sqrt{1 - (y - \theta)^2}$, per $y \in (\theta - 1, \theta + 1)$, con $\theta \in \Re$ parametro ignoto. È facile verificare, che la verosimiglianza per θ, relativa all'intera osservazione $\underline{y}$, è

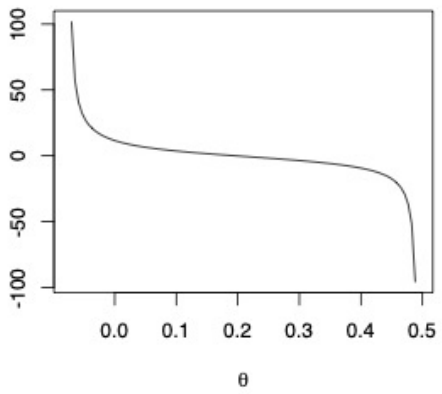
$$L(\theta) = \left\{ \begin{array}{cc} \prod_{i=1}^{n} [1 - (y_i - \theta)^2]^{1/2} & \text{se } \; y_{(n)} - 1 \leq \theta \leq y_{(1)} + 1 \\ 0 & \text{altrove} \end{array} \right\}.$$

Nella regione dello spazio parametrico Θ dove la verosimiglianza è nulla è ragionevole porre $l(\theta) = -\infty$. Quindi,

$$l(\theta) = \left\{ \begin{array}{cc} (1/2) \sum_{i=1}^{n} \log[1 - (y_i - \theta)^2] & \text{se } \; y_{(n)} - 1 \leq \theta \leq y_{(1)} + 1 \\ -\infty & \text{altrove} \end{array} \right\}.$$

Inoltre, per $\theta \in (y_{(n)} - 1, \; y_{(1)} + 1)$ la log-verosimiglianza è funzione liscia, con $l_*(\theta) = \sum_{i=1}^{n} \frac{y_i - \theta}{1 - (y_i - \theta)^2}$ e $l_{**}(\theta) < 0$. Ne consegue che la stima di massima verosimiglianza $\hat{\theta}$ può essere individuata in corrispondenza dell'unico punto di stazionarieta (se esiste) di $l_*(\theta)$, ossia tentando di risolvere l'equazione $l_*(\theta) = 0$, per $\theta \in (y_{(n)} - 1, \; y_{(1)} + 1)$.

[3] Si rifletta sulla ragionevolezza di tale scelta.



A titolo esemplificativo, la figura a fianco mostra l'andamento della derivata della log-verosimiglianza al variare di θ in $(y_{(n)}-1,\ y_{(1)}+1)$, quando $n = 10$ e l'osservazione campionaria è costituita dai valori -0,326, 0,924, -0,120, 0,887, -0,506, 0,795, 0,596, -0,354, 0,011, -0,364. Si può notare, in questo caso, l'esistenza di un'unica radice dell'equazione di verosimiglianza, che però non può essere ottenuta analiticamente.

Esercizio. Con riferimento all'Esempio 6.7, si ottenga la stima di θ fornita dal metodo dei momenti.

0.7 Alcune caratteristiche della funzione di verosimiglianza

Abbiamo definito la funzione di verosimglianza per il parametro θ che indicizza un modello statistico parametrico, scelto per descrivere la realtà complessa che ha generato l'osservazione campionaria $\underline{y}$ di cui disponiamo. Abbiamo anche introdotto la funzione di log-verosimiglianza, $l(\theta)$, la funzione $l_*(\theta)$, ottenuta derivando $l(\theta)$, e la funzione, $l_{**}(\theta)$ che si ottiene derivando due volte. Se k è la dimensione di θ, allora l_* ha dimensione k e l_{**} è una matrice $k \times k$. Chiameremo $l_*(\theta)$ **funzione punteggio (o score di verosimiglianza)**, mentre alla funzione l_{**} cambiata di segno ci riferiremo come alla **informazione osservata**, che indicheremo con $j(\theta)$. Quindi, $j(\theta) = -l_{**}(\theta)$. Ricordiamo che queste quantità, sono calcolabili se la verosimiglianza è regolare, e che anche $j(\theta)$ dipende, in generale, anche dai dati.
Abbiamo anche visto che se riparametrizziamo, cioè se passiamo dall'etichetta θ ad una nuova etichetta $\psi = g(\theta)$, con $g(\cdot)$ funzione biunivoca invertibile, vale la relazione $L^{\psi}(\psi) = L(g^{-1}(\psi))$. Quindi, se θ_1 e $\psi_1 = g(\theta_1)$ sono due etichette che si riferiscono allo stesso elemento della classe parametrica $\mathcal{F}$, si ha che $L^{\psi}(\psi_1) = L(g^{-1}(\psi_1)) = L(\theta_1)$. Questo vuol dire che, data l'osservazione campionaria $\underline{y}$, a θ_1 e ψ_1 è associato lo stesso valore di verosimiglianza, com'è ragionevole che sia: si dice che **la funzione di verosimiglianza è invariante rispetto a riparametrizzazioni**.

Esempio 7.1 In un esperimento chimico, in cui si vuole valutare l'efficacia di una possibile sostanza catalizzatrice, vengono effettuate $n + m$ prove, tutte indipendenti tra loro, in cui si dà luogo alla reazione di interesse, con e senza catalizzatore, rispettivamente. L'osservazione campionaria è costituita dalla n-upla $\underline{y} = \{y_1, y_2, \ldots, y_n\}$ e dalla m-upla $\underline{x} = \{x_1, x_2, \ldots, x_m\}$ di misure relative alla velocità delle reazioni (tempo necessario al completamento delle reazioni stesse). Si supponga sia ragionevole assumere per le variabili marginali Y_i, $i = 1, 2, \ldots, n$, e X_j, $j = 1, 2, \ldots, m$, una distribuzione esponenziale di parametri $\theta > 0$ e $\lambda > 0$, rispettivamente. Quindi $p_{Y_i}(y_i; \theta) = \theta e^{-\theta y_i}$ e $p_{X_j}(x_j; \lambda) = \lambda e^{-\lambda x_j}$, per $y_i > 0$ e $x_j > 0$.
Si focalizzi l'attenzione sul primo campione $\underline{y} = \{y_1, y_2, \ldots, y_n\}$. In base alle assunzioni fatte, la

funzione di verosimiglianza per θ, data l'osservazione $\underline{y}$, è

$$L(\theta) = \prod_{i=1}^{n} \theta e^{-\theta y_i} = \theta^n e^{-\theta \sum_{i=1}^n y_i},$$

e la log- verosimiglianza vale $l(\theta) = n\log(\theta) - \theta \sum_{i=1}^n y_i$. Si può facilmente ottenere la stima di massima verosimiglianza di θ, che risulta essere **(perché?)** $\hat{\theta} = n/\sum_{i=1}^n y_i$.
Per la distribuzione esponenziale il valore atteso è pari al reciproco del parametro θ: $E(Y_i) = 1/\theta$. Pertanto, se i ricercatori fossero interessati a valutare la durata media della reazione chimica con l'uso del catalizzatore, potremmo riparametrizzare, scegliendo come nuovo parametro $\mu = g(\theta) = 1/\theta$. La funzione inversa $g^{-1}(\mu)$ è semplicemente $g^{-1}(\mu) = 1/\mu$ e, dalla relazione richiamata sopra, avremmo

$$L^{\mu}(\mu) = L(g^{-1}(\mu)) = \frac{1}{\mu^n} e^{-\frac{1}{\mu}\sum_{i=1}^n y_i}.$$

Quindi otterremmo $l^{\mu}(\mu) = -n\log\mu - \frac{1}{\mu}\sum_{i=1}^n y_i$. La stima di massima verosimiglianza di μ sarebbe **(perché?)** $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n y_i$. Si può osservare, allora, che risulterebbe $\hat{\mu} = g(\hat{\theta})$. Ciò rappresenta la manifestazione di una proprietà generale della stima di massima verosimiglianza. **La stima di massima verosimiglianza è equivariante rispetto a riparametrizzazioni**: se θ è il parametro che indicizza il modello di partenza, per il quale abbiamo già calcolato la stima $\hat{\theta}$, e se l'interesse si rivolge ad un'altra entità, funzione (biunivoca) $\psi = g(\theta)$, allora per la stima di massima verosimiglianza $\hat{\psi}$ vale la relazione $\hat{\psi} = g(\hat{\theta})$.
Come esercizio, si ottengano la funzione di verosimiglianza per la coppia (θ, λ), data l'osservazione $(\underline{y}, \underline{x})$, calcolando la log-verosimiglianza, la funzione punteggio e l'informazione osservata.

Per dimostrare, in termini generali, la proprietà di equivarianza della stima di massima verosimiglianza, possiamo ragionare come segue. Sappiamo che $L^{\psi}(\psi) = L(g^{-1}(\psi))$; per $\hat{\theta}$, stima di massima verosimiglianza di θ, indichiamo con $\hat{\psi}$ la sua immagine attraverso la funzione $g(\cdot)$, cioè poniamo $\hat{\psi} = g(\hat{\theta})$. Allora $L(\hat{\theta}) \geq L(\theta)$ per ogni $\theta \in \Theta$, e si ha

$$L(\hat{\theta}) = L(g^{-1}(\hat{\psi})) = L^{\psi}(\hat{\psi}) \geq L(\theta) = L(g^{-1}(\psi)) = L^{\psi}(\psi),$$

da cui $L^{\psi}(\hat{\psi}) \geq L^{\psi}(\psi)$ per ogni valore $\psi \in \Psi$.

Si possono anche ricavare delle relazioni che ci permettono di vedere come cambiano alcune "quantità" di verosimiglianza quando si riparametrizza. In particolare, partendo da $L^{\psi}(\psi) = L(g^{-1}(\psi))$, si può facilmente osservare che $l^{\psi}(\psi) = l(g^{-1}(\psi))$, e, nel caso di parametro unidimensionale (e verosimiglianza regolare), derivando, si può ottenere $l^{\psi}_{*}(\psi) = l_{*}(g^{-1}(\psi))\frac{dg^{-1}(\psi)}{d\psi}$ e $l^{\psi}_{**}(\psi) = l_{**}(g^{-1}(\psi))[\frac{dg^{-1}(\psi)}{d\psi}]^2 + l_{*}(g^{-1}(\psi))\frac{d^2g^{-1}(\psi)}{d\psi^2}$.

Rimaniamo sul caso di parametro θ unidimensionale, cioè sul caso $k = 1$ (e verosimiglianza per θ, $L(\theta)$, regolare). Se indichiamo, come fatto finora, con $\hat{\theta}$ la stima di massima verosimiglianza, la funzione $L(\theta)/L(\hat{\theta})$, che chiameremo **verosimiglianza normalizzata**, è una forma equivalente per l'inferenza su θ **(perché?)**; passando al logaritmo, otteniamo la **log-verosimiglianza normalizzata** $l(\theta) - l(\hat{\theta})$. Data la regolarià della verosimiglianza, possiamo approssimare localmente la log-verosimiglianza, mediante sviluppo di Taylor del secondo ordine. Risulta che, in un intorno di $\hat{\theta}$,

vale la relazione

$$l(\theta) \approx l(\hat{\theta}) + (\theta - \hat{\theta}) l_*(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 l_{**}(\hat{\theta}) = l(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 l_{**}(\hat{\theta}) = l(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^2 j(\hat{\theta}).$$

Questa espressione mostra che, almeno localmente, in un intorno della stima di massima verosimiglianza, la log-verosimiglianza normalizzata è approssimabile con la parabola $-\frac{1}{2}(\theta - \hat{\theta})^2 j(\hat{\theta})$. Si parla qundi di **approssimazione parabolica** per la log-verosimiglianza normalizzata:

$$l(\theta) - l(\hat{\theta}) \approx -\frac{1}{2}(\theta - \hat{\theta})^2 j(\hat{\theta}).$$

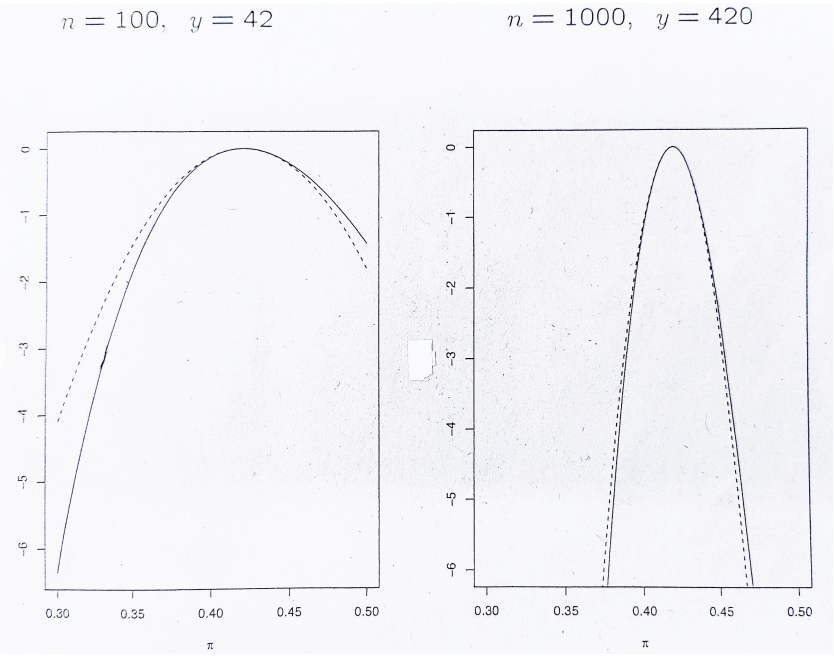
Naturalmente, l'approssimazione è tanto più accurata quanto il punto θ è vicino alla stima $\hat{\theta}$ (la parabola approssimante ha vertice in $\hat{\theta}$). Inoltre, la funzione approssimante ha derivata prima pari a $-(\theta - \hat{\theta}) j(\hat{\theta})$ e derivata seconda pari a $-j(\hat{\theta})$. Quindi, per θ fissato in un intorno di $\hat{\theta}$, più è grande l'informazione osservata calcolata in $\hat{\theta}$, maggiore è la "caduta" della log-verosimiglianza quando si passa da $\hat{\theta}$ a θ. In altri termini, maggiore è $j(\hat{\theta})$, maggiore è la capacità della log-verosimiglianza (data l'osservazione di cui si dispone) di "discriminare" tra valori diversi del parametro in un intorno del suo punto di massimo. È intuitivamente ragionevole associare a tale capacità della log-verosimiglianza la quantità di *informazione* contenuta nei dati. Possiamo perciò concludere che l'informazione osservata rappresenta una misura dell'informazione fornita dai dati.
Nel caso di parametro con dimensione $k > 1$, θ è un vettore, $j(\theta)$ una matrice e la funzione approssimante localmente la log-verosimiglianza normalizzata ha la forma di un paraboloide. Vale infatti la relazione:

$$l(\theta) - l(\hat{\theta}) \approx -\frac{1}{2}(\theta - \hat{\theta})^\top j(\hat{\theta})(\theta - \hat{\theta}).$$

Esempio 7.2 Un apparecchio elettronico è programmato per rispondere ad un impulso accendendo casualmente una lampadina verde o una lampadina rossa. La probabilità con cui la macchina accende la lampadina verde ad ogni impulso, diciamo π, è stata fissta al momento della programmazione ed è ignota agli utenti. In una serie di $n = 100$ prove si inviano altrettanti impulsi all'apparecchio che accende la lampadina verde un numero $y = 42$ di volte, in totale. In questa situazione, volendo fare inferenza su π, è naturale scegliere per la variabile Y che genera l'osservazione il modello binomiale, di indice n e parametro π. Alla luce di quanto già visto nell'esempio 2.2, la funzione di verosimiglianza per π è

$$L(\pi) = \pi^y (1 - \pi)^{n-y},$$

e la funzione di log-verosimiglianza vale $l(\pi) = y \log(\pi) + (n-y) \log(1-\pi)$. Quindi, $l_*(\pi) = \frac{y}{\pi} - \frac{n-y}{1-\pi}$ e $l_{**}(\pi) = -\frac{y}{\pi^2} - \frac{n-y}{(1-\pi)^2}$, con $l_{**}(\pi) < 0$ per ogni $\pi \in (0,1)$: l'unica radice dell'equazione $l_*(\pi) = 0$, che si ottiene quando $\pi = y/n$, è anche l'unico punto di massimo della log-verosimiglianza. Pertanto $\hat{\pi} = y/n = 0,42$ è stima di massima verosimiglianza di π. Inoltre, $j(\pi) = \frac{y}{\pi^2} + \frac{n-y}{(1-\pi)^2}$ e $j(\hat{\pi}) = \frac{n^2}{y} + \frac{n^2}{n-y} = 238,095 + 172,413 = 410,51$. Ne segue che l'approssimazione per la log-verosimiglianza normalizzata è $l(\pi) - l(\hat{\pi}) \approx -\frac{410,51}{2}(\pi - 0,42)^2$.



La figura a fianco riproduce la log-verosimiglianza normalizzata (linea continua) e l'approssimazione parabolica, sia nel caso che abbiamo esaminato, sia nel caso in cui il numero di prove nell'esperimento fosse 1000 e si osservassero 420 accensioni della lampadina verde (aumento delle informazioni nei dati).

0.8 Proprietà dello stimatore di massima verosimiglianza

Nella sezione 6 abbiamo visto come si possa calcolare, con diverse strategie a seconda dei casi, la stima di massima verosimiglianza del parametro θ che indicizza il modello statistico $\mathcal{F}$ scelto per la variabile $\underline{Y}$, di cui è realizzazionne l'osservazione campionaria $\underline{y}$. In ogni caso, alla stima $\hat{\theta}(\underline{y})$ rimane associato lo stimatore $\hat{\theta}(\underline{Y})$, le cui caratteristiche devono essere studiate. In particolare, è essenziale riuscire a rispodere a domande del tipo: lo stimatore di massima verosimiglianza è consistente? Quale è la variabilità associata alle stime che produce? Per rispondere a quesiti del genere è necessario acquisire informazioni sulla distribuzione di $\hat{\theta}(\underline{Y})$ e su come essa varia al crescere della dimensione campionaria.
Per alcuni degli esempi trattati nelle sezioni precedenti, acquisire tali informazioni è piuttosto semplice.

Nell'esempio 6.3, θ è il parametro ignoto in una distribuzione di Poisson e quindi rappresenta la media del carattere studiato (numero di clienti che frequenta la filiale bancaria al mercoledì). Abbiamo visto che lo stimatore di massima verosimiglianza è $\hat{\theta}(\underline{Y}) = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, cioè la variabile media campionaria. Richiamando la **legge dei grandi numeri** (per variabili indipendenti ed identicamente distribuite), si può immediatamente concludere che $\bar{Y} \xrightarrow{p} E(Y_i) = \theta$. Quindi $\hat{\theta}(\underline{Y}) \xrightarrow{p} \theta$ ed è stimatore consistente di θ. Inoltre, sappiamo che $E\left(\bar{Y}\right) = E(Y_i) = \theta$, che $Var\left(\bar{Y}\right) = Var(Y_i)/n$ e, per il **teorema limite centrale**, possiamo scrivere

$$\bar{Y} \dot{\sim} N\left(E(Y_i), Var(Y_i)/n\right).$$

Pertanto, nel caso di c.c.s. da una variabile di Poisson di parametro θ, lo stimatore di massima verosimiglianza di θ è stimatore consistente, non distorto ed ha varianza pari alla varianza della Poisson (cioè ancora θ) su n. Inoltre, possiamo approssimare la distribuzione di $\hat{\theta}(\underline{Y})$ con una legge

normale di media θ e varianza θ/n. Ci aspettiamo che l'approssimazione normale sia tanto più accurata quanto più è grande n.

Nell'esempio 6.4, l'osservazione campionaria è un c.c.s. da una variabile $N(\mu, \sigma^2)$. Lo stimatore di massima verosimiglinza è

$$\hat{\theta}(\underline{Y}) = \begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} \bar{Y} \\ \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2 \end{pmatrix};$$

ricordando ancora le proprietà delle variabili media campionaria e varianza campionaria, possiamo concludere che $\hat{\theta}(\underline{Y})$ è stimatore consistente di (μ, σ^2), cioè

$$\hat{\theta}(\underline{Y}) \xrightarrow{p} \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}.$$

Tuttavia, $\hat{\theta}(\underline{Y})$ non è stimatore non distorto di (μ, σ^2) **(perché?)**. Però, sappiamo che, in questo caso, la media campionaria $\bar{Y}$ ha distribuzione esatta $N(\mu, \sigma^2/n)$, e che $n\hat{\sigma}^2/\sigma^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2/\sigma^2$ ha distribuzione esatta $\chi^2(n-1)$. Inoltre, sempre sotto campionamento casuale semplice da una variabile normale, $\bar{Y}$ e $\hat{\sigma}^2$ sono indipendenti. Quindi di fatto conosciamo la distribuzione esatta dello stimatore di massima verosimiglianza: con le informazioni che abbiamo, potremmo, per esempio, calcolare la densità congiunta di $(\hat{\mu}, \hat{\sigma}^2)$.

Nell'esempio 7.1 si fa riferimento a due c.c.s. indipendenti, da due variabili esponenziali. Con riferimento al primo campione, abbiamo visto che lo stimatore di massima verosimiglianza del parametro θ è

$$\hat{\theta}(\underline{Y}) = \frac{n}{\sum_{i=1}^n Y_i} = 1/\bar{Y}.$$

Quindi, $\hat{\theta}(\underline{Y})$ è funzione **continua** di $\bar{Y}$ su $(0, +\infty)$: $\hat{\theta}(\underline{Y}) = g(\bar{Y}) = 1/\bar{Y}$. Dato che $\bar{Y} \xrightarrow{p} E(Y_i) = 1/\theta$, possiamo concludere che $\hat{\theta}(\underline{Y}) \xrightarrow{p} g(1/\theta) = \theta$. Dunque, anche in questo caso lo stimatore di massima verosimiglianza è stimatore consistente per il parametro che vuole stimare. Non è pero stimatore non distorto. Infatti la funzione $g(\cdot)$, che lega $\hat{\theta}(\underline{Y})$ alla variabile media campionaria $\bar{Y}$, è strettamente convessa su $(0, +\infty)$. Inoltre la variabile $\bar{Y}$ non è degenere; pertanto, se applichiamo la **diseguaglianza di Jensen**, risulta che

$$E[\hat{\theta}(\underline{Y})] = E[g(\bar{Y})] > g(E(\bar{Y})) = \theta.$$

Per quanto riguarda la distribuzione dello stimatore $\hat{\theta}(\underline{Y})$, possiamo ancora sfruttare il legame con $\bar{Y}$. Infatti, nel caso che stiamo trattando, $\sum_{i=1}^n Y_i$ ha distribuzione (esatta) gamma con parametro di forma n e parametro di scala θ, cioè $n\bar{Y} \sim Ga(n, \theta)$ **(perché?)**. Essendo $g(\cdot)$ funzione **continua e invertibile**, possiamo calcolare la densità di $\hat{\theta}(\underline{Y})$ come

$$p_{\hat{\theta}}(u; \theta) = p_{n\bar{Y}}(g^{-1}(u)) \left| \frac{dg^{-1}(u)}{du} \right|, \quad \text{per} \quad u > 0,$$

dove $p_{n\bar{Y}}(\cdot)$ è la funzione di densità della $Ga(n, \theta)$.

Nell'esempio 6.2 la funzione di verosimiglianza non è regolare. In ogni caso abbiamo individuato lo stimatore di massima verosimglianza che risulta essere $\hat{\theta}(\underline{Y}) = Y_{(n)} = max\{Y_1, Y_2, \ldots, Y_n\}$. Per stabilire le caratteristiche di $\hat{\theta}(\underline{Y})$, in questo caso conviene partire dal cercare di individuarne la distribuzione.
Possiamo calcolare la sua funzione di ripartizione, osservando che il supporto di $\hat{\theta}(\underline{Y})$ è l'intervallo $(0, \theta]$. Sappiamo che

$$F_{\hat{\theta}}(u) = F_{Y_{(n)}}(u) = \Pr\{Y_{(n)} \leq u\} = \Pr\{Y_1 \leq u\} \Pr\{Y_2 \leq u\} \cdots \Pr\{Y_n \leq u\};$$

quindi, $F_{\hat{\theta}}(u) = (\frac{u}{\theta})^n$ se $0 < u \leq \theta$ e $F_{\hat{\theta}}(u) = \{^{0 \quad \text{se} \quad u \leq 0}_{1 \quad \text{se} \quad u > \theta}\}$. Pertanto, si ha che, comunque si fissi $\epsilon > 0$, $\Pr\{\hat{\theta}(\underline{Y}) > \theta - \epsilon\} = 1 - (\frac{\theta - \epsilon}{\theta})^n$, e questa quantità converge a 1 quando $n \to +\infty$. Ne segue che $\hat{\theta}(\underline{Y}) \xrightarrow{p} \theta$ **(si riveda la definizione di convergenza in probabilità)**, ed è quindi stimatore consistente di θ. È però stimatore distorto. Infatti vale la relazione $\hat{\theta}(\underline{Y}) < \theta$ con probabilità 1, quindi risulta necessariamente $E(\hat{\theta}(\underline{Y})) < \theta$.

Esercizio. Si stabilisca se lo stimatore scelto nell'Esempio 6.6 è stimatore consistente del parametro θ che indicizza il modello.

Come abbiamo visto, nelle situazioni a cui si riferiscono gli esempi sopra considerati (e certamente anche in altre particolari situazioni), con strumenti e ragionamenti *ad hoc*, si riesce a stabilire che lo stimatore di massima verosimiglianza è stimatore consistente per il parametro che indicizza il modello parametrico fissato, e si riescono a recuperare le informazioni necessarie sulla sua distribuzione esatta o su qualche forma di approssimazione[4]. Detto questo, è chiaro che, dal punto di vista pratico, sarebbe molto utile poter disporre di un risultato generale sul comportamento dello stimatore di massima verosimiglianza. Fortunatamente un tale risultato esiste ed è utilizzabile quando sono soddisfatte certe condizioni, che chiameremo **condizioni di regolarità**.

Se $\mathcal{F}$ è il modello parametrico (identificabile e correttamente specificato) scelto per l'osservazione $\underline{y}$, realizzazione della variabile $\underline{Y}$, con spazio campionario $\mathcal{Y}$ e parametro $\theta \in \Theta \subseteq \Re^k$, diciamo che ci troviamo nel contesto di **un problema regolare di stima** o, più generalmente, che siamo **sotto condizioni di regolarità**, se:

- $\mathcal{F}$ ha verosimiglianza regolare;
- lo spazio campionario non dipende dal parametro (tutti gli elementi di $\mathcal{F}$ hanno lo stesso supporto);
- si possono scambiare le operazioni di integrazione rispetto a $\underline{y}$ (o suoi elementi) e derivazione rispetto a θ (o suoi elementi), nelle espressioni matematiche che coinvolgono tali operazioni;
- esistono (finiti) i valori attesi delle derivate (eventualmente parziali), fino al terzo ordine, della log-verosimiglianza.

[4] Gli esempi discussi mostrano anche un risultato da tenere presente: lo stimatore di massima verosimiglianza può essere distorto.

L'ultimo punto non deve sorprendere perché, abbiamo già detto che la funzione di verosimiglianza (come le altre quantità di verosimiglianza) dipende anche dai dati e, che quando occorre, $L(\theta)$ conviene scriverla nella forma estesa $L(\theta; \underline{y})$. Dato che $\underline{y}$ è realizzazione di una variabile casuale, tale sarà anche $L(\theta; \underline{y})$, e con $L(\theta; \underline{Y})$ indicheremo l'oggetto casuale corrispondente. Quindi non deve meravigliare l'idea di poter prendere in considerazione caratteristiche, come i valori attesi, ad esempio, delle **quantità di verosimiglianza**: è chiaro che, nel caso, si fa riferimento agli oggetti casuali associati.
Prendiamo ad esempio lo score si verosimiglianza $l_*(\theta)$. La variabile casuale corrispondente è $l_*(\theta; \underline{Y})$. Chiediamoci: sotto $\mathcal{F}$, quante distribuzioni possiamo considerare per tale variabile? Naturalmente, la risposta è ...tante quanti sono gli elementi di $\mathcal{F}$, perché la distribuzione di $l_*(\theta; \underline{Y})$ dipende dalla distribuzione che consideriamo per $\underline{Y}$, quindi dall'etichetta che scegliamo. Potremo perciò avere la distribuzione di $l_*(\theta; \underline{Y})$, quando la legge di $\underline{Y}$ è quella etichettata da un certo θ_1, o da θ_2, e così via, e ci riferiremo a quelle distribuzioni come alle distribuzioni **sotto** θ_1, sotto θ_2, eccetera. Dal punto di vista notazionale, con riferimento al valore atteso, per esempio, indicheremo con $E_{\theta_1}(l_*(\theta; \underline{Y}))$ il valore atteso della distribuzione di $l_*(\theta; \underline{Y})$ sotto θ_1. È importante osservare che, in questo caso, l'etichetta che caratterizza la legge è diversa da quella presente come argomento nell'espressione dello score (θ_1 da una parte e θ dall'altra). Quando le due etichette coincidono scriviamo $E_\theta(l_*(\theta; \underline{Y}))$ o, per semplificare la notazione stessa, usiamo la scrittura $E(l_*(\theta; \underline{Y}))$. Questa convenzione che semplifica la notazione sarà ricorrente nel seguito.

Chiariti questi aspetti, possaimo dare i risultati generali che riguardano il comportamento dello stiamtore di massima verosimiglianza sotto condizioni di regolarità.
Sotto condizioni di regolarità,

- **lo stimatore di massima verosimiglianza $\hat{\theta}(\underline{Y})$ è stimatore consistente del parametro θ che indicizza la classe $\mathcal{F}$; in altri termini, $\hat{\theta}(\underline{Y}) \xrightarrow{p} \theta$, sotto θ, per ogni $\theta \in \Theta$;**
- **$\hat{\theta}(\underline{Y})$ è asintoticamente normale, e per la sua legge vale l'approssimazione**

$$\hat{\theta}(\underline{Y}) \dot{\sim} N_k(\theta, i^{-1}(\theta))$$

sotto θ, per ogni $\theta \in \Theta$.

Nell'espressione sopra, N_k (il pedice è omesso nel caso $k = 1$) indica la legge normale di dimensione k e $i(\theta)$ è l'**informazione attesa**, definita come valore atteso dell'informazione osservata, $i(\theta) = E(j(\theta; \underline{Y})) = E(-l{*}{*}(\theta; \underline{Y}))$.

Il secondo risultato, quello distributivo, è essenziale perché consente, per esempio, di associare alla stima di massima verosimiglianza una misura della sua variabilità: $i^{-1}(\theta)$ rappresenta un'approssimazione della matrice di varianza-covarianza dello stimatore di massima verosimiglianza.

Nell'esempio 6.4, in cui abbiamo trattato il caso di osservazione campionaria costituita da un c.c.s. da $N(\mu, \sigma^2)$, posto $\theta = (\mu, \sigma^2)$, abbiamo visto che

$$l_{**}(\theta) = \begin{pmatrix} \frac{\partial^2 l(\theta)}{\partial \mu^2} & \frac{\partial^2 l(\theta)}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 l(\theta)}{\partial \mu \partial \sigma^2} & \frac{\partial^2 l(\theta)}{\partial (\sigma^2)^2} \end{pmatrix} = \begin{pmatrix} -\frac{n}{\sigma^2} & \frac{n\mu}{\sigma^4} - \frac{\sum_{i=1}^n y_i}{\sigma^4} \\ \frac{n\mu}{\sigma^4} - \frac{\sum_{i=1}^n y_i}{\sigma^4} & \frac{n}{2\sigma^4} - \frac{\sum_{i=1}^n (y_i - \mu)^2}{\sigma^6} \end{pmatrix}.$$

Allora, dato che $E_\theta\left[-\frac{n\mu}{\sigma^4}+\frac{\sum_{i=1}^n y_i}{\sigma^4}\right]=0$ e $E_\theta\left[-\frac{n}{2\sigma^4}+\frac{\sum_{i=1}^n (y_i-\mu)^2}{\sigma^6}\right]=\frac{n}{2\sigma^4}$, risulta

$$i(\theta)=\begin{pmatrix}\frac{n}{\hat\sigma^2} & 0\\ 0 & \frac{n}{2\hat\sigma^4}\end{pmatrix},$$

e per $\hat\theta=(\hat\mu,\hat\sigma^2)$ vale l'approssimazione normale bidimensionale, con vettore delle medie $\theta=(\mu,\sigma^2)$ e matrice di varianza-covarianza

$$i^{-1}(\theta)=\begin{pmatrix}\frac{\sigma^2}{n} & 0\\ 0 & \frac{2\sigma^4}{n}\end{pmatrix}.$$

Sotto condizioni di regolarità, $i^{-1}(\theta)$ può essere stimata (consistentemente) da $i^{-1}(\hat\theta)$ o da $j^{-1}(\hat\theta)$. Inoltre, il risultato che sancisce la possibilità di approssimare la legge di $\hat\theta(\underline{Y})$ con la distribuzione normale continua a valere anche con $i^{-1}(\theta)$ sostituita da una sua stima. Potremo cioè scrivere

$$\hat\theta(\underline{Y})\dot\sim N_k(\theta,i^{-1}(\hat\theta)) \quad \text{o} \quad \hat\theta(\underline{Y})\dot\sim N_k(\theta,j^{-1}(\hat\theta)),$$

e questa cosa ci permetterà di costruire, attorno allo stimatore di massima verosimiglianza, quantità (approssimativamente) pivotali e statistiche test. Nella prossima sezione vedremo una dimostrazione parziale del risultato riguardante la normalità asintotica, nel caso più semplice in cui $k=1$.

Molti dei modelli parametrici di uso comune soddisfano le condizioni di regolarità. Ciò rende i risultati dati sopra abbastanza generali e parecchio utili in pratica. Se ripensiamo agli esempi trattati finora, le condizioni sono soddisfatte in tutti i casi, a parte i casi degli esempi 6.2 e 6.6 che coinvolgono la distribuzione uniforme, e dell'esempio 6.7. Quindi per tutti gli esempi discussi, a parte tre, avremmo potuto immediatamente dedurre la consistenza dello stimatore di massima verosimiglianza (per il parametro che indicizza il modello) e avremmo potuto fornire con facilità un'approssimazione per la sua distribuzione. Per quanto riguarda quest'ultimo aspetto, ricordiamoci che, quando usiamo l'approssimazione normale non abbiamo idea, tipicamente, della sua accuratezza come surrogato della vera e ignota distribuzione. Quello che sappiamo è che, in generale, il livello di accuratezza cresce al crescere dell'informazione, quindi della dimensione campionaria, ma dipende anche dalla dimensione del parametro. Come abbiamo visto con gli esempi, a volte è possibile ottenere, con relativo impegno, la distribuzione esatta di $\hat\theta(\underline{Y})$; allora, nei casi di scarsa informazione fornita dai dati (dimensione campionaria relativamente piccola), è consigliabile basare su di essa il processo inferenziale.[5]

0.9 Proprietà di $\hat\theta(\underline{Y})$: alcuni aspetti teorici

Prima di procedere occorre richiamare un paio di importanti proprietà dello score di verosimiglianza. Queste proprietà sono note come **prima e seconda identità di Bartlett**, rispettivamente. **Sotto condizioni di regolarità, valgono le seguenti relazioni:**

[5] Solo per completezza e a titolo informativo, esistono metodi studiati per aumentare l'accuratezza dell'approssimazione normale per la distribuzione di $\hat\theta(\underline{Y})$. Tali argomenti però vanno oltre gli obiettivi che motivano questi "Appunti".

- $E(l_*(\theta;\underline{Y}))=0$, **sotto** θ, **per ogni** $\theta\in\Theta$;
- $Var(l_*(\theta;\underline{Y}))=E(-l{*}{*}(\theta;\underline{Y}))=i(\theta)$, **sotto** θ, **per ogni** $\theta\in\Theta$.

Le due relazioni, sanciscono che, sotto condizioni di regolarità, lo score di verosimiglianza (inteso come oggetto casuale), $l_*(\theta;\underline{Y})$, ha come valore atteso il vettore nullo e come matrice di varianza-covarianza la matrice di informazione attesa, sotto θ.[6]

Ora consideriamo il caso in cui l'osservazione campionaria sia costituita da un c.c.s., e il parametro che indicizza il modello fissato abbia dimensione $k=1$. Sia $p_{Y_i}(y_i;\theta)$ la legge (funzione di probabilità o densità) della generica variabile Y_i coinvolta. La funzione di verosimiglianza per θ, data l'intera ossevazione $\underline{y}$, è il prodotto di n contributi, se n è la dimensione del campione,

$$L(\theta;\underline{y})=\prod_{i=1}^{n}p_{Y_i}(y_i;\theta),$$

dove il generico contributo $p_{Y_i}(y_i;\theta)$ arriva dalla generica i-esima unità campionaria. Passando al logaritmo, possiamo notare che le log-verosimiglianza è la somma di n contributi,

$$l(\theta;\underline{y})=\sum_{i=1}^{n}\log[p_{Y_i}(y_i;\theta)],$$

e lo stesso vale per lo score $l_*(\theta;\underline{y})$ e per $l_{**}(\theta;\underline{y})$. Più precisamente, abbiamo che

$$l_*(\theta;\underline{y})=\sum_{i=1}^{n}\frac{d}{d\theta}\log[p_{Y_i}(y_i;\theta)]\quad\text{e}\quad l_{**}(\theta;\underline{y})=\sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log[p_{Y_i}(y_i;\theta)].$$

Se pensiamo alle variabili casuali associate a tali quantità, possiamo scrivere

$$l_*(\theta;\underline{Y})=\sum_{i=1}^{n}V_i\quad\text{e}\quad l_{**}(\theta;\underline{Y})=\sum_{i=1}^{n}Q_i,$$

con $V_i=\frac{d}{d\theta}\log[p_{Y_i}(Y_i;\theta)]$ e $Q_i=\frac{d^2}{d\theta^2}\log[p_{Y_i}(Y_i;\theta)]$. Quindi, $l_*(\theta;\underline{Y})$ e $l_{**}(\theta;\underline{Y})$ sono somme di variabili indipendenti ed identicamente distribuite **(perché?)**. Ora, dalla prima identità di Bartlett, deriva che necessariamente $E(V_i)=0$, per ogni i, sotto θ **(perché?)**. Inoltre, dalla seconda identità di Bartlett, ricaviamo che $Var(V_i)=E(-Q_i)=i(\theta)/n$, per ogni i, sotto θ, per ogni $\theta\in\Theta$ **(perché?)**. Ne risulta che, per il teorema limite centrale e la legge dei grandi numeri, rispettivamente,

$$\sqrt{n}\,\frac{l_*(\theta;\underline{Y})}{n}\xrightarrow{d}N(0,i_1(\theta))\quad\text{e}\quad\frac{-l_{**}(\theta;\underline{Y})}{n}\xrightarrow{p}i_1(\theta),$$

sotto θ, avendo posto $i_1(\theta)=i(\theta)/n$.

[6] Se è chiaro quanto detto nella sezione precedente, si comprenderà che, per esempio, $E_{\theta_1}(l_*(\theta;\underline{Y}))\neq 0$, in generale.

Supponiamo ora di aver dimostrato la consistenza di $\hat{\theta}(\underline{Y})$.[7] L'obiettivo qui è fornire una bozza di dismostrazione della sua normalità asintotica. Per fare ciò, partiamo dal fatto che (si può dimostrare) vale la seguente approssimazione per lo score di verosimiglianza, sotto θ:

$$\sqrt{n} l_*(\hat{\theta}; \underline{Y}) \doteq \sqrt{n} l_*(\theta; \underline{Y}) + \sqrt{n}(\hat{\theta} - \theta) l_{**}(\theta; \underline{Y}).$$

Il simbolo $\doteq$ indica equivalenza in probabilità tra i due membri (di destra e sinistra), quando n tende a $+\infty$. Sotto θ, questa approssimazione, che deriva da uno sviluppo di Taylor al primo ordine, è possibile perché siamo sotto condizioni di regolarità e perché abbiamo assunto la consistenza di $\hat{\theta}$, che quindi, per n sufficientemete grande, "sarà prossimo a θ". Il membro di sinistra vale 0 **(perché?)**. Quindi possiamo scrivere che

$$\sqrt{n}(\hat{\theta} - \theta) \doteq \frac{\sqrt{n} l_*(\theta; \underline{Y})}{j(\theta; \underline{Y})} = \frac{\sqrt{n} l_*(\theta; \underline{Y})/n}{j(\theta; \underline{Y})/n},$$

sotto θ. Dato che

$$\sqrt{n}\, \frac{l_*(\theta; \underline{Y})}{n} \xrightarrow{d} N(0, i_1(\theta)) \quad \text{e} \quad \frac{j(\theta; \underline{Y})}{n} \xrightarrow{p} i_1(\theta),$$

possiamo concludere che, sotto θ,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, i_1^{-1}(\theta)),$$

e questo giustifica la scrittura $\hat{\theta}(\underline{Y}) \dot{\sim} N(\theta, i^{-1}(\theta))$.

Esempio 9.1 Riconsideriamo il problema trattato nell'esempio 2.4. I dati $\underline{y} = \{y_1, y_2, \ldots, y_n\}$ sono un c.c.s. da una varaiabile Poisson di parametro θ. La funzione di verosimiglianza per θ è $L(\theta) = \theta^{\sum_{i=1}^n y_i} e^{-n\theta}$, e la log-verosimiglianza vale $l(\theta) = \log(\theta) \sum_{i=1}^n y_i - n\theta$. Vogliamo verificare la vaidità delle due identità di Bartlet. Derivando la log-verosimiglianza otteniamo la funzione punteggio

$$l_*(\theta; \underline{Y}) = \frac{\sum_{i=1}^n Y_i}{\theta} - n,$$

che ora consideriamo come oggetto casuale. Derivando ancora e cambiando di segno, otteniamo l'informazione osservata

$$j(\theta; \underline{Y}) = -l_{**}(\theta; \underline{Y}) = \frac{\sum_{i=1}^n Y_i}{\theta^2}.$$

Ora, per ogni $\theta \in \Theta = (0, +\infty)$, si ha che

$$E_\theta[l_*(\theta; \underline{Y})] = \frac{\sum_{i=1}^n E_\theta(Y_i)}{\theta} - n = \frac{n\theta}{\theta} - n = 0,$$

e

$$i(\theta) = E_\theta[j(\theta; \underline{Y})] = \frac{\sum_{i=1}^n E_\theta(Y_i)}{\theta^2} = \frac{n\theta}{\theta^2} = \frac{n}{\theta}.$$

Inoltre,

$$Var_\theta[l_*(\theta; \underline{Y})] = \frac{\sum_{i=1}^n Var_\theta(Y_i)}{\theta^2} = \frac{n}{\theta}.$$

Quindi, come ci si aspettava, abbiamo che, sotto θ, $E[l_*(\theta; \underline{Y})] = 0$ e $Var[l_*(\theta; \underline{Y})] = i(\theta)$. Infine, in base ai risultati forniti nella sezione precedente, possiamo affermare che, sotto θ, $\hat{\theta}(\underline{Y}) = \frac{\sum_{i=1}^n Y_i}{n}$ è stimatore consistente del parametro θ e che vale l'approssimazione $\hat{\theta}(\underline{Y}) \dot{\sim} N(\theta, \theta/n)$ (e anche $\hat{\theta}(\underline{Y}) \dot{\sim} N(\theta, \hat{\theta}/n)$).

[7] Per questa dimostrazione, che lasciamo ai corsi avanzati, gioca un ruolo essenziale la prima identià di Bartlett.

0.10 Quantità pivotali e statistiche test basate su $\hat{\theta}(\underline{Y})$ ($k = 1$)

Sia θ il parametro undimensionale che indicizza il modello $\mathcal{F}$ scelto per l'osservazione campionaria $\underline{y} = \{y_1, y_2, \ldots, y_n\}$, realizzazione della variabile $\underline{Y}$.
Supponiamo di essere nel contesto di un problema regolare di stima e sia $\hat{\theta}(\underline{Y})$ lo stimatore di massima verosimiglianza per θ. Allora vale l'approssimazione **(perché?)**

$$\frac{\hat{\theta}(\underline{Y}) - \theta}{\sqrt{j^{-1}(\hat{\theta})}} \dot{\sim} N(0,1),$$

sotto θ. Posto $r_e(\theta; \underline{Y}) = \frac{\hat{\theta}(\underline{Y}) - \theta}{\sqrt{j^{-1}(\hat{\theta})}}$, la quantià $r_e(\theta; \underline{Y})$ è una funzione dei dati (o delle variabili di cui i dati sono realizzazioni) e del parametro θ, la cui distribuzione è approssimativamente nota nella forma e non dipende dal parametro né da altri elementi ignoti. Essa è, quindi, **quantità approssimativamente pivotale** per l'inferenza su θ.
La quantità $r_e(\theta; \underline{Y})$, che chiameremo **quantità di tipo Wald**, può essere allora utilizzata per costruire **intervalli di confidenza** (approssimati) per θ. Infatti, scelto un livello di confideza nominale $1 - \alpha$, con $\alpha \in (0, 1)$, possiamo scrivere:

$$\Pr_\theta\{|r_e(\theta; \underline{Y})| \leq z_{1-\alpha/2}\} \approx 1 - \alpha,$$

se $z_{1-\alpha/2}$ è il quantile di ordine $1 - \alpha/2$ della normale standard, ovvero se $\Pr\{Z \leq z_{1-\alpha/2}\} = 1 - \alpha/2$, quando Z indica una variabile casuale normale standard. Dunque,

$$\begin{aligned}
\Pr_\theta\{|r_e(\theta; \underline{Y})| \leq z_{1-\alpha/2}\} &= \Pr_\theta\left\{-z_{1-\alpha/2} \leq \frac{\hat{\theta}(\underline{Y}) - \theta}{\sqrt{j^{-1}(\hat{\theta})}} \leq z_{1-\alpha/2}\right\} \\
&= \Pr_\theta\left\{\hat{\theta}(\underline{Y}) - z_{1-\alpha/2}\sqrt{j^{-1}(\hat{\theta})} \leq \theta \leq \hat{\theta}(\underline{Y}) + z_{1-\alpha/2}\sqrt{j^{-1}(\hat{\theta})}\right\} \\
&\approx 1 - \alpha,
\end{aligned}$$

e l'intervallo $\hat{\Theta}(\underline{Y}) = \left\{\theta \in \Theta : \ \hat{\theta}(\underline{Y}) - z_{1-\alpha/2}\sqrt{j^{-1}(\hat{\theta})} \leq \theta \leq \hat{\theta}(\underline{Y}) + z_{1-\alpha/2}\sqrt{j^{-1}(\hat{\theta})}\right\}$ costituisce un intervallo (casuale) di confidenza per θ, con livello di confidenza nominale $1 - \alpha$. Il livello di confidenza reale dell'intervallo non è noto. Ci aspettiamo che sia via via più vicino a quello nominale al crescescere della dimensione del campione. In altri termini, possiamo dire che, il livello nominale approssima quello reale ignoto. La realizzazione dell'intervallo casuale la otteniamo calcolando gli estremi dell'intervallo stesso, usando i dati effettivamente disponibili $\{y_1, y_2, \ldots, y_n\}$.
Altre quantità approssimativamente pivotali per l'inferenza su θ, basate sempre sullo stimatore di massima verosimiglianza e simili nella forma a quella che abbiamo indicato con $r_e(\theta; \underline{Y})$, sono $\frac{\hat{\theta}(\underline{Y}) - \theta}{\sqrt{i^{-1}(\hat{\theta})}}$, $\frac{\hat{\theta}(\underline{Y}) - \theta}{\sqrt{j^{-1}(\theta)}}$ e $\frac{\hat{\theta}(\underline{Y}) - \theta}{\sqrt{i^{-1}(\theta)}}$, per le quali vale lo stesso risultato, in termini di distribuzione approssimante sotto θ. La quantità $r_e(\theta; \underline{Y})$ è quella generalmente più usata per ragioni di praticità. Per esempio, l'uso della quantità $\frac{\hat{\theta}(\underline{Y}) - \theta}{\sqrt{j^{-1}(\theta)}}$ non porta, in generale, a calcolare esplicitamente l'intervallo (o più in generale la **regione**) di confidenza; spesso si deve ricorrere a metodi numerici per

individuare l'insieme

$$\hat{\Theta}(\underline{y}) = \left\{ \theta \in \Theta : \ -z_{1-\alpha/2} \leq \frac{\hat{\theta}(\underline{y}) - \theta}{\sqrt{j^{-1}(\theta)}} \leq z_{1-\alpha/2} \right\}.$$

Esempio 10.1 Sia $\underline{y} = \{y_1, y_2, \ldots, y_n\}$, un c.c.s. da una distribuzione gamma $Ga(2, \theta)$; quindi la generica variabile Y_i ha funzione di densità $p_{Y_i}(y_i; \theta) = \theta^2 y_i e^{-\theta y_i}$, per $y_i > 0$ e $\theta > 0$. Il modello associato all'intera osservazione $\underline{y}$ è indicizzato dal parametro θ, ed ha funzione di verosimiglianza $L(\theta) = \prod_{i=1}^n \theta^2 y_i e^{-\theta y_i} \propto \theta^{2n} e^{-\theta \sum_{i=1}^n y_i}$. Ne segue che la funzione di log-verosimiglianza è $l(\theta) = 2n \log(\theta) - \theta \sum_{i=1}^n y_i$ e che $l_*(\theta) = 2n/\theta - \sum_{i=1}^n y_i$ e $l_{**}(\theta) = -2n/\theta^2$. Essendo $l_{**}(\theta) < 0$ per ogni $\theta \in \Theta = (0, +\infty)$, l'unico punto di stazionarietà, che troviamo come radice dell'equazione $l_*(\theta) = 0$, è anche l'unico punto di massimo della log-verosimiglianza. Quindi, risulta $\hat{\theta} = 2n / \sum_{i=1}^n y_i$. Se supponiamo che sia $n = 10$ e che l'osservazione campionaria sia tale per cui $\sum_{i=1}^{10} y_i = 6,5$, otteniamo che la stima di massima verosimiglianza vale $\hat{\theta} = 20/6,5 = 3,077$ e che $j(\hat{\theta}) = 2,11$. Pertanto un intervallo di confidenza per θ, di livello nominale 0,95, ha per estremi $\hat{\theta} - z_{0,975}\sqrt{j^{-1}(\hat{\theta})} = 1,727$ e $\hat{\theta} + z_{0,975}\sqrt{j^{-1}(\hat{\theta})} = 4,426$, essendo $z_{0,975} = 1,96$. Si osservi che in questo caso informazione osservata e informazione attesa coincidono **(perché?)**. Quindi se utilizzassimo la quantità $\frac{\hat{\theta}(\underline{Y}) - \theta}{\sqrt{i^{-1}(\hat{\theta})}}$, otterremmo lo stesso intervallo.

Una considerazione importante è la seguente. Vale la relazione

$$\Pr_\theta\{|r_e(\theta; \underline{Y})| \leq z_{1-\alpha/2}\} = \Pr_\theta\{[r_e(\theta; \underline{Y})]^2 \leq z^2_{1-\alpha/2}\} \approx 1 - \alpha,$$

e, se poniamo $W_e(\theta; \underline{Y}) = [r_e(\theta; \underline{Y})]^2$, risulta che, sotto θ, $W_e(\theta; \underline{Y})$ ha distribuzione approssimabile con quella $\chi^2(1)$ **(perché?)**, e $z^2_{1-\alpha/2} = \chi^2_{1-\alpha}(1)$; quindi, l'intervallo di confidenza $\hat{\Theta}(\underline{Y}) = \left\{ \theta \in \Theta : \ \hat{\theta}(\underline{Y}) - z_{1-\alpha/2}\sqrt{j^{-1}(\hat{\theta})} \leq \theta \leq \hat{\theta}(\underline{Y}) + z_{1-\alpha/2}\sqrt{j^{-1}(\hat{\theta})} \right\}$, basato su $r_e(\theta; \underline{Y})$, è ottenibile anche come

$$\hat{\Theta}(\underline{Y}) = \left\{ \theta \in \Theta : W_e(\theta; \underline{Y}) \leq \chi^2_{1-\alpha}(1) \right\}.$$

Qui $\chi^2_{1-\alpha}(1)$ indica il quantile di ordine $1 - \alpha$ della distribuzione $\chi^2(1)$, cioè della distribuzione chi-quadrato con un grado di libertà.

Supponiamo ora di dover verificare l'ipotesi $H_0 : \theta = \theta_0$, con θ_0 valore fissato. All'ipotesi H_0 (ipotesi nulla), può essere contrapposta un'alternativa unilaterale o bilaterale. In ogni caso, una **statistica test** adeguata può essere ottenuta facilmente, specificando opportunamente la quantià $r_e(\theta; \underline{Y})$. Infatti, se in tale quantità poniamo $\theta = \theta_0$, otteniamo una funzione solo dei dati (o delle variabili da cui i dati sono generati), quindi una statistica, che è $r_e(\theta_0; \underline{Y}) = \frac{\hat{\theta}(\underline{Y}) - \theta_0}{\sqrt{j^{-1}(\hat{\theta})}}$. Sotto H_0, $r_e(\theta_0; \underline{Y})$ ha distribuzione approssimabile con la legge normale standard. Inoltre, $r_e(\theta_0; \underline{y})$ è una misura della discrepanza tra quanto osservato (l'osservazione è espressa in sintesi da $\hat{\theta}$), e quanto previsto da H_0; tale misura è data secondo una metrica opportuna, basata direttamente su $\hat{\theta}$. La forma di $r_e(\theta_0; \underline{y})$ ne rende semplice e intuitivo l'utilizzo, che dipende dall'alternativa, H_1, contrapposta ad H_0. Se, per esempio, l'alternativa fosse unilaterale destra, $H_1 : \theta > \theta_0$, si sarebbe indotti

naturalmente a rifiutare l'ipotesi nulla a fronte di valori *grandi* della statistica test. Per alternative bilaterali, invece, indicherebbero *contrarietà* dei dati rispetto ad H_0, valori grandi in modulo di $r_e(\theta_0;\underline{y})$. Visto che **sotto** H_0 la distribuzione della statistica test è approssimabile con quella normale standard, l'uso della calibrazione di $r_e(\theta_0;\underline{y})$, mediante i quantili di questa distribuzione notevole, rende facilmente calcolabili la **regione di rifiuto** o la **regione di accettazione** di **livello di significatività** nominale (fissato) α.

Esempio 10.2 Riconsideriamo la situazione trattata nell'esempio precedente (Esempio 10.1) e supponiamo di voler risolvere il **problema di verifica d'ipotesi** caratterizzato dal **sistema d'ipotesi** $H_0 : \theta = \theta_0 = 2$ contro $H_1 : \theta > \theta_0$. Il valore osservato della statistica test risulterebbe in questo caso $r_e^{oss} = r_e(\theta_0;\underline{y}) = r_e(2;\underline{y}) = \frac{3{,}077-2}{\sqrt{1/2{,}11}} = 1{,}565$. Essendo l'ipotesi alternativa unilaterale destra, si rifiuta H_0 per valori grandi e, in particolare, la regione critica di livello di significatività nominale α ha la forma $R_\alpha = (z_{1-\alpha}, +\infty)$, con $z_{1-\alpha}$ quantile di ordine $1-\alpha$ della distribuzione normale standard. Con $\alpha = 0{,}05$, $z_{1-\alpha} = 1{,}64$ e $r_e^{oss} \notin R_{0{,}05}$. Pertanto, alla luce dei dati osservati, non si rifiuterebbe l'ipotesi nulla, ad un livello di significatività nominale del 5%. Volendo calcolare il **livello di significatività osservato**, α_{oss}, avremmo $\alpha_{oss} = \Pr_{H_0}\{r_e(\theta_0;\underline{Y}) \geq r_e^{oss}\} \approx \Pr\{Z > r_e^{oss}\} = \Pr\{Z > 1{,}565\}$, dove Z è una variabile casuale con distribuzione normale standard.
Se l'alternativa fosse bilaterale, $H_1 : \theta \neq \theta_0$, la regione di accettazione di livello nominale $\alpha = 0{,}05$ risulterebbe $A_{0{,}05} = (-z_{0{,}975}, +z_{0{,}975}) = (-1{,}96, +1{,}96)$. Dato che $r_e^{oss} \in A_{0{,}05}$, anche in questo caso, con i dati di cui si dispone, non si potrebbe rifiutare l'ipotesi nulla ad un livello di significatività nominale del 5%. Volendo calcolare il livello di significatività osservato in questo secondo caso, avremmo $\alpha_{oss} = 2\min\{\Pr_{H_0}\{r_e(\theta_0;\underline{Y}) \geq r_e^{oss}\},\ \Pr_{H_0}\{r_e(\theta_0;\underline{Y}) \leq r_e^{oss}\}\} \approx 2\Pr\{Z > |r_e^{oss}|\} = 2\Pr\{Z > 1{,}565\}$.

A questo punto possiamo porci la seguente domada: è possibile risolvere problemi di verifica d'ipotesi, su un parametro unidimensionale, utilizzando una statistica test derivata da $W_e(\theta;\underline{y})$? La risposta è affermativa, purché si faccia attenzione ad un elemento importante. L'ipotesi alternativa, contrapposta all'ipotesi nulla H_0, rappresenta, in un problema di verfica d'ipotesi su un parametro scalare, la(e) direzione(i) degli scostamenti da H_0 che è(sono) di interesse nel problema stesso. La statistica test $r_e(\theta_0;\underline{y})$ contiene, per sua natura, l'informazione (proveniente, ovviamente, dai dati) relativa alla direzione dello scostamento tra realtà e quanto fissato da H_0: essa, in effetti, può assumere sia valori positivi sia valori negativi. Al contrario, la statistica $W_e(\theta_0;\underline{y})$, ottenibile in maniera ovvia dalla quantità $W_e(\theta;\underline{y})$, perde parte dell'informazione, essendo pari al quadrato di $r_e(\theta_0;\underline{y})$. Essa assume solo valori positivi, e può essere utilizzata **solo** per risolvere problemi di verifica d'ipotesi con alternativa bilaterale. Più precisamente, se il sistema d'ipotesi fissato fosse: $H_0 : \theta = \theta_0$ contro $H_1 : \theta \neq \theta_0$, l'ipotesi nulla sarebbe rifiutata per valori *grandi* di $W_e(\theta_0;\underline{y})$. Dato che, **sotto** H_0, $W_e(\theta_0;\underline{y})$ è realizzazione di una variabile casuale con distribuzione approssimabile dalla legge $\chi^2(1)$, la regione critica del test, di livello di significatività nominale α, ha la forma $R_\alpha = (\chi^2_{1-\alpha}(1), +\infty)$. Dovrebbe essere chiaro, inoltre, che, in pratica, per un problema di verifica d'ipotesi sul parameto unidimensionale θ con alternativa bilaterale, l'uso della statistica test $r_e(\theta_0;\underline{y})$ e l'uso della statistica $W_e(\theta_0;\underline{y})$ sono assolutamente equivalenti.

Esempio 10.3 Facciamo riferimento alla seconda parte dell'esempio precedente (Esempio 10.2), e

consideriamo l'alternativa bilaterale $H_1 : \theta \neq \theta_0$, contrapposta all'ipotesi nulla $H_0 : \theta = \theta_0 = 2$. Essendo $r_e^{oss} = 1,565$, risulta $We^{oss} = W_e(2; \underline{y}) = 2,45$. Con il livello nominale α fissato pari a 0,05, la regione di rifiuto del test basato su $W_e(\theta_0; \underline{y})$ risulta essere $R_{0,05} = (\chi^2_{0,95}(1), +\infty) = (3,84, +\infty)$ e, con i dati osservati non si può quindi rifiutare l'ipotesi nulla, almeno ad un livello di significatività nominale del 5%. Per ottenere il livello di significatività osservato, α_{oss}, occorre calcolare la $\Pr_{H_0}\{W_e(\theta_0; \underline{Y}) \geq W_e^{oss}\} \approx \Pr\{V \geq W_e^{oss}\}$, con V variabile casuale con distribuzione $\chi^2(1)$.

Osservazione Finora abbiamo visto come usare le quantità $r_e(\theta; \underline{y})$ e $W_e(\theta; \underline{y})$ per costruire intervalli di confidenza per il parametro θ che indicizza il modello statistico parametrico $\mathcal{F}$ considerato, oppure per derivare statistiche test per la soluzione di problemi di verifica d'ipotesi su θ. C'è da dire che, in questo secondo caso, cioè quando il problema da risolvere è un problema di verifica d'ipotesi, quasi sempre si preferisce usare la statistiche test $\frac{\hat{\theta}(\underline{Y})-\theta_0}{\sqrt{j^{-1}(\theta_0)}}$ (o eventualmente la statistica $\frac{\hat{\theta}(\underline{Y})-\theta_0}{\sqrt{i^{-1}(\theta_0)}}$) o la corrispondente versione al quadrato, che sono asintoticamente equivalenti a $r_e(\theta_0; \underline{Y})$ e $W_e(\theta_0; \underline{Y})$, rispettivamente. Tale scelta è motivata dall'idea che l'uso di una stima della varianza di $\hat{\theta}(\underline{Y})$ specifica sotto l'assunto che H_0 sia vera, renda lo strumento statistico (quindi il test) più accurato.

Osservazione Le quantità come $r_e(\theta; \underline{Y})$ e $W_e(\theta; \underline{Y})$, e le statistiche test da esse derivate, sono basate direttamente sullo stimatore di massima verosimiglianza $\hat{\theta}(\underline{Y})$. I risultati generali su $\hat{\theta}(\underline{Y})$, in particolare quello distributivo (approssimazione normale), che valgono se sono soddisfatte le condizioni di regolarità, ci agevolano molto nell'uso di tali strumenti per la costruzione di intervalli di confidenza per θ e per la soluzione di problemi di verifica d'ipotesi. Il prezzo che dobbiamo pagare per questa "agevolazione" sta nel fatto che dobbiamo accontentarci di risultati "approssimati". Se la dimensione campionaria è sufficientemente grande, ci aspettiamo (o speriamo in) un errore, dovuto all'approssimazione e che non possiamo controllare, relativamente piccolo. Se però siamo costretti a lavorare con campioni di piccola taglia, l'errore che commettiamo quando usiamo una distribuzione approssimante in luogo della vera e ignota distribuzione di $\hat{\theta}(\underline{Y})$ può essere rilevante. In questi casi, per liberarsi dell'errore di approssimazione e fare inferenza "esatta" bisogna riuscire a costruire quantità esattamente pivotali (o statistiche test la cui distribuzione sia esattamente nota sotto H_0) e quindi, in ultima analisi, studiare la distribuzione di $\hat{\theta}(\underline{Y})$ o di sue funzioni. Come mostra il prossimo esempio, in alcune situazioni, almeno, questo processo è possibile e può essere anche poco laborioso.

Esempio 10.4 Continuiamo ancora ad analizzare la situazione trattata in tutti gli esempi di questa sezione: $\underline{y} = \{y_1, y_2, \ldots, y_n\}$, c.c.s. da una distribuzione $Ga(2, \theta)$. Abbiamo visto che lo stimatore di massima verosimiglianza per θ è $\hat{\theta}(\underline{Y}) = 2n/\sum_{i=1}^n Y_i$. Essendo $Y_i \sim Ga(2; \theta)$, si ha che $\sum_{i=1}^n Y_i \sim Ga(2n; \theta)$ e $\theta \sum_{i=1}^n Y_i \sim Ga(2n; 1)$ sotto θ (**perché?**). Infine, dato che $\sum_{i=1}^n Y_i = 2n/\hat{\theta}(\underline{Y})$, risulta che $2n\theta/\hat{\theta}(\underline{Y}) \sim Ga(2n; 1)$ sotto θ, e $T(\theta; \hat{\theta}(\underline{Y})) = 4n\theta/\hat{\theta}(\underline{Y}) \sim Ga(4n/2; 1/2)$. Quindi, $T(\theta; \hat{\theta}(\underline{Y})) \sim \chi^2(4n)$ sotto θ, ed è una quantità esattamente pivotale per l'inferenza su θ, basata sullo stimatore di massima verosimiglianza. Possiamo allora scrivere che

$$\Pr_\theta \left\{ \chi^2_{\alpha/2}(4n) \leq 4n\theta/\hat{\theta}(\underline{Y}) \leq \chi^2_{1-\alpha/2}(4n) \right\} = 1 - \alpha,$$

per ogni $\alpha \in (0,1)$, da cui

$$\Pr_\theta \left\{ \chi^2_{\alpha/2}(4n)\hat{\theta}(\underline{Y})/4n \leq \theta \leq \chi^2_{1-\alpha/2}(4n)\hat{\theta}(\underline{Y})/4n \right\} = 1-\alpha.$$

In definitiva, un intervallo di confidenza di livello **esatto** $1-\alpha$ per θ ha per estremi $\hat{\theta}(\underline{y})\chi^2_{\alpha/2}(4n)/4n$ e $\hat{\theta}(\underline{y})\chi^2_{1-\alpha/2}(4n)/4n$. Con la sintesi dell'osservazione fornita nell'esempio 10.1, cioè $n = 10$ e $\sum_{i=1}^{10} y_i = 6,5$, l'intervallo per θ, di livello esatto 0,95, risulta essere $(3,077 \times 24,43/40,\ 3,077 \times 59,34/40) = (1,88, 4,56)$.
Supponiamo ora di voler risolvere il problema di verifica d'ipotesi considerato nell'esempio 10.2: $H_0 : \theta = \theta_0 = 2$ contro $H_1 : \theta > \theta_0$. Dalla quantità pivotale $T(\theta; \hat{\theta}(\underline{Y}))$ possiamo ricavare la statistica test $T(\theta_0; \hat{\theta}(\underline{Y})) = 4n\theta_0/\hat{\theta}(\underline{Y})$, che sotto H_0 ha distribuzione esatta $\chi^2(4n)$. Dato che l'ipotesi alternativa fissata è unilaterale destra, considerata la natura della statistica test, sono i valori *piccoli* della statistica stessa ad essere contrari all'ipotesi nulla **(perché?)**. Pertanto, si rifiuta H_0 ad un livello di significatività esatto α se $T^{oss} = 4n\theta_0/\hat{\theta}(\underline{y})$ appartiene alla regione critica $R_\alpha = (0, \chi^2_\alpha(4n))$. Con la sintesi dell'osservazione campionaria di cui disponiamo, risulta essere $T^{oss} = 80/3,077 = 26$, e la regione critica di livello 0,05 è $R_{0,05} = (0,\ 24,43)$; quindi, ad un livello di significatività **esatto** del 5%, l'ipotesi nulla non può essere rifiutata, alla luce dell'informazione disponibile. Per quanto riguarda il livello di significatività osservato, esso si ottiene come la $\Pr_{H_0}\{T(\theta_0; \hat{\theta}(\underline{Y})) \leq T^{oss}\} = \Pr\{V \leq T^{oss}\}$, con V variabile casuale con distribuzione $\chi^2(40)$.

0.11 Quantità pivotali e statistiche test basate su $l_*(\theta; \underline{Y})$ ($k = 1$)

Dai risultati visti nel Paragrafo 9, quando valgono le condizioni di regolarità, possiamo scrivere che

$$\frac{l_*(\theta; \underline{Y})}{\sqrt{i(\theta)}} \dot{\sim} N(0,1),$$

sotto θ. Questo significa che la quantità $r_u(\theta; \underline{Y}) = l_*(\theta; \underline{Y})/\sqrt{i(\theta)}$ è approssimativamente pivotale per l'inferenza su θ e può essere usata per costruire intervalli (e più in generale regioni) di confidenza per il parametro. In effetti, dato che vale la relazione

$$\begin{aligned} \Pr_\theta\{|r_u(\theta; \underline{Y})| \leq z_{1-\alpha/2}\} &= \Pr_\theta \left\{ \left| \frac{l_*(\theta; \underline{Y})}{\sqrt{i(\theta)}} \right| \leq z_{1-\alpha/2} \right\} \\ &\approx 1-\alpha, \end{aligned}$$

l'insieme dei punti dello spazio parametrico

$$\hat{\Theta}(\underline{y}) = \left\{ \theta \in \Theta : -z_{1-\alpha/2} \leq r_u(\theta; \underline{y}) \leq z_{1-\alpha/2} \right\},$$

costituisce una regione di confidenza (nei casi più semplici un untervallo) per θ, di livello di confidenza nominale $1-\alpha$. Il livello di confidenza reale non sarà ovviamente noto; ci aspettiamo, in generale, che sia vicino a quello nominale quando la dimensione campionaria è elevata. La quantità $r_u(\theta; \underline{y})$ è una quantità **di tipo score**, visto che è basata sulla funzione punteggio. Essa è in realtà poco

utilizzata perché poco comoda: in effetti, a parte casi particolari, l'ottenimento della regione di confidenza di livello nominale voluto comporta la verifica della relazione $-z_{1-\alpha/2} \leq r_u(\theta; \underline{y}) \leq z_{1-\alpha/2}$ (per i punti dello spazio parametrico) per via numerica. Come per la quantità $r_e(\theta; \underline{Y})$, anche in questo caso si può considerarne il quadrato $W_u(\theta; \underline{Y}) = [r_u(\theta; \underline{Y})]^2$, che ha distribuzione approssimabile con la legge $\chi^2(1)$, sotto θ. Pertanto, le regioni di confidenza fornite da $r_u(\theta; \underline{Y})$ sono anche ottenibili come

$$\hat{\Theta}(\underline{y}) = \left\{\theta \in \Theta : W_u(\theta; \underline{y}) \leq \chi^2_{1-\alpha}(1)\right\}.$$

Supponiamo ora di dover verificare l'ipotesi $H_0 : \theta = \theta_0$, con θ_0 valore fissato. Al solito, all'ipotesi nulla H_0, può essere contrapposta un'alternativa unilaterale o bilaterale. In ogni caso, una statistica test adeguata può essere ottenuta facilmente dalla quantià $r_u(\theta; \underline{Y})$. Infatti, se in tale quantità poniamo $\theta = \theta_0$, otteniamo una funzione solo dei dati (o delle variabili da cui i dati sono generati), quindi una statistica, che è $r_u(\theta_0; \underline{y}) = \frac{l_*(\theta_0; \underline{y})}{\sqrt{i(\theta_0)}}$. Sotto H_0, $r_u(\theta_0; \underline{Y})$ ha distribuzione approssimabile con la legge normale standard. Inoltre, $r_u(\theta_0; \underline{y})$ è una misura della discrepanza tra quanto osservato (l'osservazione è espressa in sintesi da $l_*(\hat{\theta}; \underline{y}) = 0$), e quanto previsto da H_0; tale misura è data secondo una metrica opportuna, basata sulla funzione punteggio. La forma di $r_u(\theta_0; \underline{y})$ ne rende semplice l'utilizzo, che dipende dall'alternativa, H_1, contrapposta ad H_0. Se, per esempio, l'alternativa fosse unilaterale destra, $H_1 : \theta > \theta_0$, si sarebbe indotti naturalmente a rifiutare l'ipotesi nulla a fronte di valori *grandi* della statistica test[8] . Per alternative bilaterali, invece, indicherebbero *contrarietà* dei dati rispetto ad H_0, valori grandi in modulo di $r_u(\theta_0; \underline{y})$. Visto che sotto H_0 la distribuzione della statistica test è approssimabile con quella normale standard, l'uso dei quantili di questa distribuzione notevole, rende facilmente calcolabili la regione di rifiuto o la regione di accettazione di livello di significatività nominale (fissato) α.
Come avviene per $W_e(\theta_0; \underline{y})$, anche $W_u(\theta_0; \underline{y})$ può essere utilizzata come statistica test per risolvere problemi di verifica d'ipotesi, ma solo quando l'ipotesi alternativa è bilaterale. La distribuzione approssimante la legge di $W_u(\theta_0; \underline{Y})$, sotto H_0, è di nuovo la distribuzione $\chi^2(1)$. La regione critica del test, di livello di significatività nominale α, ha la forma $R_\alpha = (\chi^2_{1-\alpha}(1), +\infty)$.

0.12 Quantità pivotali e statistiche test basate sul log-rapporto di verosimiglianza $W(\theta; \underline{Y})$ $(k = 1)$

Riconsideriamo il problema della costruzione di un intervallo (o regione) di confidenza per il parametro unidimensionale θ che indicizza il modello statistico parametrico $\mathcal{F}$ scelto per l'osservazione campionaria (c.c.s). Abbiamo visto che, se ricorriamo all'uso della quantità pivotale $W_e(\theta)$, l'intervallo di livello nominale $1 - \alpha$ è costituito da tutti i punti dello spazio parametrico per i quali $W_e(\theta; \underline{Y}) \leq \chi^2_{1-\alpha}(1)$. Formalmente, cioè,

$$\hat{\Theta}(\underline{y}) = \left\{\theta \in \Theta : W_e(\theta; \underline{y}) \leq \chi^2_{1-\alpha}(1)\right\},$$

[8] $l_*(\theta; \underline{y})$ è funzione tipicamente decrescente, almeno localmente, in un intorno di $\hat{\theta}$.

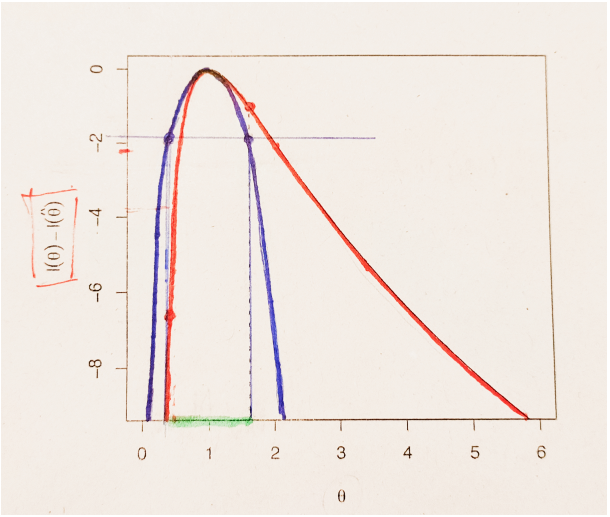
ovvero,

$$\hat{\Theta}(\underline{y}) = \left\{\theta \in \Theta : -\frac{1}{2}W_e(\theta;\underline{y}) \geq -\frac{1}{2}\chi^2_{1-\alpha}(1)\right\}.$$

Ricordando che $-\frac{1}{2}W_e(\theta;\underline{y}) = -\frac{1}{2}r^2_e(\theta;\underline{y})$ rappresenta la funzione (parabola) appossimante la log-verosimiglianza normalizzata $l(\theta;\underline{y}) - l(\hat{\theta};\underline{y})$ in un intorno della stima $\hat{\theta}$, possiamo dedurre che l'intervallo costruito mediante $W_e(\theta;\underline{y})$ (o equivalentemente mediante $r_e(\theta;\underline{y})$) raccoglie i punti dello spazio parametrico Θ per i quali la parabola **approssimante** la log-verosimiglianza normalizzata ha un valore superiore ad una soglia fissata. Tali punti non sono però quelli effettivamente a log-verorosimiglianza più elevata, soprattutto se l'approssimazione parabolica risulta poco accurata. In generale, una regione di confifenza per θ è un insieme che raccoglie, alla luce dell'osservazione campionaria, i punti dello spazio parametrico più **plausibili** come valore dell'ignoto parametro, secondo un qualche criterio di plausibilità. Se usiamo la verosimiglianza come misura di plausibilità dei vari elementi di Θ, è del tutto coerente l'idea di costruire regioni di confidenza per θ che rispettino esattamente tale criterio, non solo approssimativamente.

Esempio 12.1 Supponiamo di disporre di un c.c.s. $\underline{y} = \{y_1, y_2, \ldots, y_n\}$, di dimensione n=10, da una variabile esponenziale di media ignota θ. La funzione di verosimiglianza per θ vale $L(\theta) = \theta^{-n}e^{-n\bar{y}/\theta}$ e la log-verosimiglianza risulta essere $l(\theta) = -n\log\theta - n\bar{y}/\theta$, essendo $\bar{y}$ la media campionaria. Inoltre, $l_*(\theta) = -n/\theta + n\bar{y}/\theta^2$ e $l_{**}(\theta) = n/\theta^2 - 2n\bar{y}/\theta^3$. Ne deriva che $\hat{\theta} = \bar{y}$, $l(\hat{\theta}) = -n\log\hat{\theta} - n$ e $j(\hat{\theta}) = n/\hat{\theta}^2$.
Se assumiamo che l'osservazione campionaria sia espressa in sintesi da $\bar{y} = 1$, possiamo ricavare $W_e(\theta) = 10(\theta-1)^2$ e $-\frac{1}{2}W_e(\theta) = -5(\theta-1)^2$. Inoltre, $l(\theta) - l(\hat{\theta}) = -10(\log\theta + 1/\theta - 1)$. Un intervallo di confidenza per la media θ, di livello nominale 0,95 e basato su W_e sarà dato da

$$\hat{\Theta}(\underline{y}) = \left\{\theta \in \Theta : -\frac{1}{2}W_e(\theta;\underline{y}) \geq -\frac{1}{2}\chi^2_{1-\alpha}(1)\right\} = \left\{\theta \in \Theta : -5(\theta-1)^2 \geq -1.92\right\}.$$



La figura a fianco riproduce la log-verosimiglianza normalizzata (linea rossa), l'approssimazione parabolica e l'intervallo di confidenza di livello nominale 0,95 per θ, basato su $W_e(\theta)$. Tale intervallo risulta essere (0,38, 1,62): si può osservare che, per esempio, il valore 1,7 non è contenuto nell'intervallo, pur essendo più plausibile (a log-verosimiglianza più elevata) del valore 0,39.

Per ottenere una regione di confidenza per θ costituita effettivamente solo da punti dello spazio parametrico a più elevata verosimiglianza, potremmo raccogliere tutti i punti $\theta \in \Theta$ tali che $l(\theta) -$

$l(\hat{\theta}) \geq h$, con la soglia h scelta opportunamente. Come scegliamo h? Dato che, sotto θ,

$$l(\theta; \underline{Y}) - l(\hat{\theta}; \underline{Y}) \doteq -\frac{1}{2} W_e(\theta; \underline{Y}) \quad \text{e} \quad 2[l(\hat{\theta}; \underline{Y}) - l(\theta; \underline{Y})] \doteq W_e(\theta; \underline{Y}),$$

possiamo concludere che $2[l(\hat{\theta}; \underline{Y}) - l(\theta; \underline{Y})] \dot{\sim} \chi^2(1)$, sotto θ. Pertanto, la regione cercata è data da

$$\hat{\Theta}(\underline{y}) = \left\{ \theta \in \Theta : l(\theta; \underline{y}) - l(\hat{\theta}; \underline{y}) \geq h \right\} = \left\{ \theta \in \Theta : 2[l(\hat{\theta}; \underline{y}) - l(\theta; \underline{y})] \leq \chi^2_{1-\alpha}(1) \right\},$$

se $1 - \alpha$ è il livello nominale fissato. La quantità $W(\theta; \underline{Y}) = 2[l(\hat{\theta}; \underline{Y}) - l(\theta; \underline{Y})]$ è detta **log-rapporto di verosimiglianza**. Si ossevi che la regione sopra indicata può anche essere ottenuta come $\hat{\Theta}(\underline{y}) = \left\{ \theta \in \Theta : l(\theta; \underline{y}) \geq l(\hat{\theta}; \underline{y}) - \frac{1}{2}\chi^2_{1-\alpha}(1) \right\}$. Ne caso sell'Esempio 12.1, la regione di confidenza per θ, basata sul log-rapporto di verosimiglianza $W(\theta; \underline{y})$, di livello nominale 0,95, è data da

$$\hat{\Theta}(\underline{y}) = \left\{ \theta \in \Theta : l(\theta; \underline{y}) - l(\hat{\theta}; \underline{y}) \geq -\frac{1}{2}\chi^2_{1-\alpha}(1) \right\} = \{ \theta \in \Theta : -10(\log\theta + 1/\theta - 1) \geq -1,92 \}.$$

Osservando la figura nella pagina precedente, si nota che tale regione è un intervallo: a differenza di quello che si ottiene usando $W_e(\theta; \underline{y})$, non si tratta però di un intervallo centrato su $\hat{\theta}$.

La quantità $W(\theta; \underline{Y})$, approssimativamente pivotale per l'inferenza su θ, può essere facilmente trasformata in una statistica test, se il problema da risolvere è un problema di verifica d'ipotesi. In effetti, se consideriamo il sistema d'ipotesi $H_0 : \theta = \theta_0$ contro $H_1 : \theta \neq \theta_0$, la statistica $W(\theta_0; \underline{Y}) = 2[l(\hat{\theta}; \underline{Y}) - l(\theta_0; \underline{Y})]$ ha distribuzione approssimante $\chi^2(1)$, sotto H_0, e valuta la discrepanza tra quanto si osserva e quanto fissato dall'ipotesi nulla direttamente in termini di log-verosimiglianza. Il valore osservato $W^{oss} = W(\theta_0; \underline{y})$ è dunque contrario ad H_0 se è sufficientemente grande: la regione critica del test basato sul log-rapporto di verosimiglianza è la stessa di quella del test basato su W_e o W_u ed è ottenibile facendo riferimento ai quantili della distribuzione $\chi^2(1)$. In particolare, per un livello di significatività nominale α, la regione critica è $R_\alpha = (\chi^2_{1-\alpha}(1), +\infty)$.

Nel caso di problemi con ipotesi alternativa unilaterale, la statistica $W(\theta_0; \underline{y})$ non può essere utilizzata. Si può però facilmente ricavare da essa uno strumento alternativo. In effetti, in analogia con $r_e(\theta_0; \underline{Y})$, per cui vale la relazione $r_e(\theta_0; \underline{Y}) = sgn(\hat{\theta} - \theta_0)\sqrt{W_e(\theta_0; \underline{Y})}$, con $sgn(\cdot)$ **funzione segno**, possiamo definire $r(\theta_0; \underline{Y}) = sgn(\hat{\theta} - \theta_0)\sqrt{W(\theta_0; \underline{Y})}$, e per tale statistica vale l'approssimazione

$$r(\theta_0; \underline{Y}) \dot{\sim} N(0, 1),$$

sotto H_0. Ne risulta che, per esempio, se è $H_1 : \theta > \theta_0$, la regione critica del test basato sul log-rapporto di verosimiglianza di livello nominale α è $R_\alpha = (z_{1-\alpha}, +\infty)$, dove, al solito, $z_{1-\alpha}$ indica il quantile di ordine $1-\alpha$ della normale standard. La statistica $r(\theta_0; \underline{Y})$ è detta **radice con segno**.

Naturalmente, la quantità $r(\theta; \underline{Y})$, per cui vale l'approssimazione $r(\theta; \underline{Y}) \dot{\sim} N(0, 1)$ sotto θ, è approssimativamente pivotale per l'inferenza su θ e può essere utilizzata per ottenere regioni di confidenza per il parametro. Tali regioni (eventualmente intervalli) coincideranno, a parità di livello di confidenza nominale, con quelle ottenibili da $W(\theta; \underline{Y})$.

Esercizio Con riferimento all'Esempio 12.1, si risolva il problema di verifica d'ipotesi $H_0 : \theta = \theta_0 = 2,5$ contro $H_1 : \theta > \theta_0$, ad un livello nominale del 5%, usando sia $r_e(\theta_0)$ sia $r(\theta_0)$. Si calcoli il livello di significatività osservato nei due casi.

0.13 Alcune considerazioni

Le tre classi di quantità di verosimiglianza (di tipo Wald, di tipo score e di tipo log-rapporto di verosimiglianza) introdotte finora producono strumenti che sono, sotto condizioni di regolarità, asintoticamente equivalenti, nel senso che

$$r_e(\theta;\underline{Y}) \doteq r_u(\theta;\underline{Y}) \doteq r(\theta;\underline{Y}) \quad \text{e} \quad W_e(\theta;\underline{Y}) \doteq W_u(\theta;\underline{Y}) \doteq W(\theta;\underline{Y}),$$

sotto θ. Ciò significa che, per grandi dimensioni campionarie, le procedure inferenziali derivanti da esse tendono a produrre risultati molto simili. Le caratteristiche e le proprietà di tali quantità sono però tipicamente diverse, così come il loro comportamento nel finito. Ciò le porta a produrre, specie con campioni di piccole dimensioni, risultati inferenziali che possono differire notevolmente anche in termini di accuratezza.

Delle tre classi considerate, le due più utilizzate sono la prima e la terza. Le quantità di tipo Wald hanno il loro principale elemento di forza nella loro semplicità, e il loro uso risulta più facilmente comprensibile agli utenti finali. Si pensi al problema della costruzione di una regione di confidenza per il parametro (unidimensionale) ignoto θ. Se si procede usando $r_e(\theta;\underline{y})$, non si incontrano grossi problemi computazionali, si ottiene sempre un intervallo e gli estremi sono calcolabili analiticamente, risultando pari a $\hat{\theta} - z_{1-\alpha/2}\sqrt{\hat{var}(\hat{\theta})}$ e $\hat{\theta} + z_{1-\alpha/2}\sqrt{\hat{var}(\hat{\theta})}$; cioè: stima del parametro più e meno l'**errore standard** (stima dello scarto quadratico medio associato alla stima del parametro) moltiplicato per una opportuna costante (spesso posta grossolanamente pari a 2 **(perché?)**). Il prezzo che si paga per questa semplicità d'uso risulta nella presenza di un vincolo (di simmetria) predeterminato sulla forma degli intervalli, vincolo spesso non giustificato alla luce dell'osservazione campionaria, in una generale scarsa accuratezza (livello di confidenza reale non prossimo a quello nominale) in campioni di dimensione medio-piccola e nel fatto che gli intervalli possono non essere (automaticamente) coerenti con la forma dello spazio parametrico.

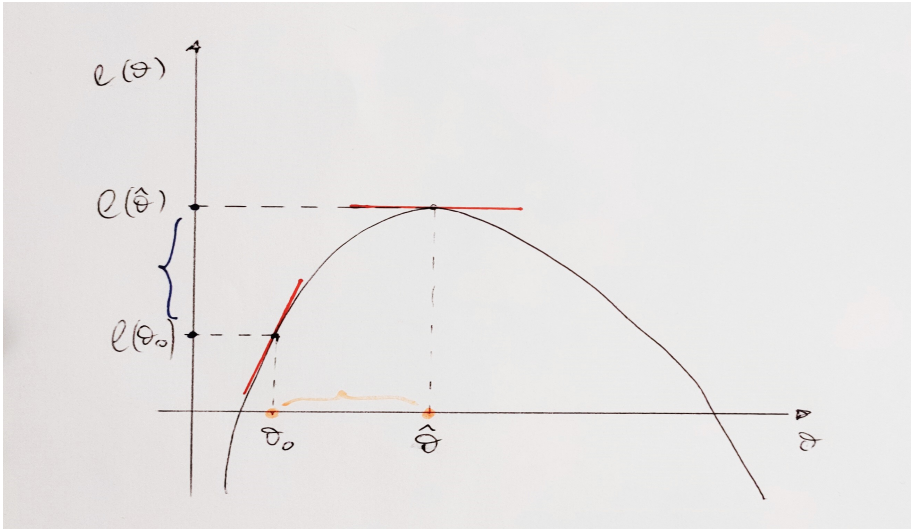
Esempio 13.1 Si supponga di osservare 2 successi in una sequenza di n=20 prove di Bernoulli, con probabilità di successo ignota θ. In questo caso, lo spazio parametrico è l'intervallo $\Theta = (0,1)$ e, con $y = 2$, risulta $l(\theta) = 2\log\theta + 18\log(1-\theta)$. La stima di massima verosimiglianza è $\hat{\theta} = y/n = 0,1$ e si ha $j^{-1}(\hat{\theta}) = \hat{\theta}(1-\hat{\theta})/n = 0,09/20 = 0,0045$. Pertanto, l'intervallo di confidenza per θ, di livello nominale 0,95 risulta essere $(0,1 - 1,96\sqrt{0,0045}, 0,1 + 1,96\sqrt{0,0045}) = (-0,031, 0,23)$. È evidente che tale intervallo non rispetta lo spazio parametrico. Occorre porre rimedio e fornire, come risultato sensato, l'intervallo (0, 0,23). Come esercizio, si ottenga l'intervallo per θ nel caso in cui $n = 40$ e $y = 4$.

La terza classe di quantità, di tipo log-rapporto di verosimiglianza, è quella che produce procedure inferenziali generalmente più affidabili. Più precisamente, per campioni di dimensioni medio piccole,

le procedure risultano più accurate (studi di simulazione hanno mostrato una generale migliore qualità dell'approssimazione della distribuzione asintotica a quella vera) e rispettano automaticamente la natura dello spazio parametrico. Questo significa che, per esempio, le regioni di confidenza costruite attraverso $r(\theta; \underline{y})$ o $W(\theta; \underline{y})$ sono sempre interne allo spazio parametrico e possono quindi essere indicate, formalmente, come $\hat{\Theta}(\underline{y}) = \{\theta : W(\theta) \leq \chi^2_{1-\alpha}(1)\}$, se ci si riferisce all'uso di W. Inoltre, tali regioni non presentano vincoli di forma predeterminati, avendo forma risultante solo dalla natura dell'osservazione $\underline{y}$ e dal modello scelto. Giusto come illustrazione, per il problema dell'Esempio 13.1, la regione di confidenza per θ di livello nominale 0,95 basato su $W(\theta)$ (o $r(\theta)$) risulta essere l'intervallo (0,017, 0,277). Naturalmente, rispetto alle quantità di tipo Wald, le quantità derivanti dal log-rapporto di verosimiglianza sono più complicate da usare, in particolare se si guarda all'aspetto computazionale.

Le quantità di tipo score sono quelle meno utilizzate in pratica perché il loro uso è più complicato (in genere) se confrontato con l'uso delle quantità di tipo Wald e non hanno le buone caratteristiche delle quantità direttamente basate sulla log-verosimiglianza.

Un'ultima considerazione riguarda le logiche su cui si basano le quantità di verosimiglianza di cui parliamo in questo paragrafo. Tutte effettuano un confronto tra una opportuna sintesi dell'osservazione campionaria e i punti dello spazio parametrico (un valore θ_0 nel caso di problema di verifica d'ipotesi). Tale confronto è fatto usando metriche opportune e, come elemento di sintesi dell'osservazione, sostanzialmente la stima $\hat{\theta}$. Nel caso di quantità di tipo Wald, il confronto avviene in termini diretti tra $\hat{\theta}$ e i punti di Θ. Nel caso di quantità di tipo score, il confronto tra la sintesi e un punto dello spazio parametrico (diciamo θ_0), avviene in termini della pendenza della retta tangente alla log-verosimiglianza. Infine, nel caso delle quantità di tipo log-rapporto di verosimiglianza, il confronto tra la sintesi dell'osservazione e i punti dello spazio parametrico avviene in termini di caduta di log-verosimiglianza. La figura che segue cerca di illustrare quanto appena detto.



0.14 Parametro multidimensionale ($k > 1$): inferenza globale e inferenza parziale

Sia ora $k > 1$. Parliamo di problema di **inferenza globale** quando l'obiettivo dell'inferenza è l'ottenimento di una regione di confidenza per l'intero parametro θ, oppure quando si tratta di risolvere un problema di verifica d'ipotesi su θ con ipotesi nulla semplice, cioè con un'ipotesi nulla che identifica uno e un solo elemento della classe parametrica $\mathcal{F}$ scelta per l'osservazione $\underline{y}$. In tutte le altre possibili situazioni, affrontiamo problemi che definiamo di **inferenza parziale**.

Nel caso di problemi di inferenza globale, le quantità più diffusamente utilizzate[9] -e di cui discuteremo- sono

$$W_e(\theta) = (\theta - \hat{\theta})^\top j(\hat{\theta})(\theta - \hat{\theta}) \quad \text{e} \quad W(\theta) = 2[l(\hat{\theta}) - l(\theta)],$$

che rappresentano le estensioni al caso multidimensionale delle quantità viste con $k = 1$ e che, sotto θ e sotto condizioni di regolarità, hanno distribuzione approssimabile con la legge $\chi^2(k)$ (tale risultato deriva dalla distribuzione asintotica normale -di dimensione k- dello stimatore di massima verosimiglianza $\hat{\theta}(\underline{Y})$, e dall'equivalenza asintotica tra $W_e(\theta; \underline{Y})$ e $W(\theta; \underline{Y})$).

Per quanto riguarda l'inferenza parziale, introdurremo nel seguito alcuni strumenti specifici adatti a trattare problemi relativi alla verifica di ipotesi o alla costruzione di regioni di confidenza per **un singolo elemento** del vettore θ, o per **una sua parte** (cioè un insieme di suoi elementi). Infine tratteremo, in termini generali, la situazione in cui il problema da risolvere è un problema di verifica d'ipotesi, in cui l'ipotesi di interesse pone fissati vincoli sui k elementi di θ.

0.15 Inferenza globale mediante $W_e(\theta; \underline{y})$

Sia $\underline{y} = \{y_1, y_2, \ldots, y_n\}$ l'osservazione campionaria e sia $\mathcal{F}$ il modello scelto per la variabile $\underline{Y}$, di cui è realizzazione $\underline{y}$. Sia $\theta \in \Theta$ il parametro che indicizza il modello e sia $k > 1$ la dimensione di θ (scriveremo anche $k = \dim(\Theta)$).

Se sono soddisfatte le condizioni di regolarità, $W_e(\theta; \underline{Y})$ è quantità approssimativamente pivotale per l'inferenza su θ, con distribuzione approssimante $\chi^2(k)$, sotto θ. Quindi

$$\Pr_\theta\{W_e(\theta; \underline{Y}) \leq \chi^2_{1-\alpha}(k)\} \approx 1 - \alpha,$$

e

$$\hat{\Theta}(\underline{Y}) = \left\{\theta \in \Theta : W_e(\theta; \underline{Y}) \leq \chi^2_{1-\alpha}(k)\right\}$$

costituisce una regione di confidenza per il parametro θ, di livello nominale $1 - \alpha$. Dato che $-\frac{1}{2}W_e(\theta; \underline{y})$ è il paraboloide che approssima localmente la log-verosimiglianza normalizzata, le regioni del tipo $\left\{\theta \in \Theta : W_e(\theta; \underline{y}) \leq \chi^2_{1-\alpha}(k)\right\}$ hanno sempre (per ogni livello di confidenza fissato) forma ellittica se $k = 2$, o , più in generale la forma di un ellissoide. Inoltre, tali regioni sono sempre centrate sulla stima $\hat{\theta}$.

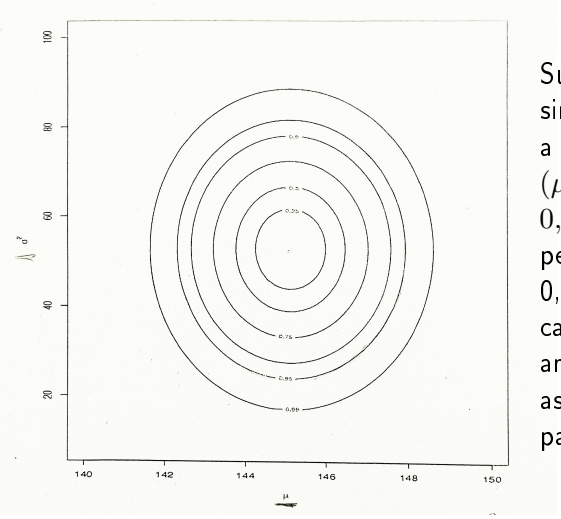
[9] La quantità $W_u(\theta) = l_*(\theta)^\top i^{-1}(\theta) l_*(\theta)$ è raramente utilizzata.

Esempio 15.1 Sia $\underline{y} = \{y_1, y_2, \ldots, y_n\}$ un c.c.s. da una distribuzione $N(\mu, \sigma^2)$. Sia $\theta = (\mu, \sigma^2)$ il parametro che indicizza il modello fissato per $\underline{Y}$. Abbiamo visto che la stima di massima verosimiglianza di θ è $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$, con $\hat{\mu}$ e $\hat{\sigma}^2$ rispettivamente media e varianza campionarie. Inoltre, abbiamo anche visto che

$$l_{**}(\hat{\theta}) = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2\hat{\sigma}^4} \end{pmatrix}$$

e, di conseguenza, risulta

$$\begin{aligned} W_e(\theta) = (\hat{\theta} - \theta)^\top j(\hat{\theta})(\hat{\theta} - \theta) &= \begin{pmatrix} \hat{\mu} - \mu \\ \hat{\sigma}^2 - \sigma^2 \end{pmatrix}^\top \begin{pmatrix} \frac{n}{\hat{\sigma}^2} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4} \end{pmatrix} \begin{pmatrix} \hat{\mu} - \mu \\ \hat{\sigma}^2 - \sigma^2 \end{pmatrix} \\ &= \frac{n}{\hat{\sigma}^2}(\hat{\mu} - \mu)^2 + \frac{n}{2\hat{\sigma}^4}(\hat{\sigma}^2 - \sigma^2)^2. \end{aligned}$$



Supponendo $n = 40$ e l'informazione espressa in sintesi da $\hat{\mu} = \bar{y} = 145,13$ e $\hat{\sigma}^2 = 7,26^2$, il grafico a fianco fornisce le regioni di confidenza per $\theta = (\mu, \sigma^2)$ basate su $W_e(\theta; \underline{y}) = 0,758(\mu - 145,13)^2 + 0,00719(\sigma^2 - 52,74)^2$ e l'approssimazione $\chi^2(2)$, per diversi livelli di confidenza nominali (0,25, 0,50, 0,75, 0,90, 0,95, 0,99). Come si può notare, a causa della forma della matrice $j(\hat{\theta})$ (ricordiamo anche che $\hat{\mu}$ e $\hat{\sigma}^2$ sono stimatori indipendenti) gli assi delle ellissi che definiscono le regioni risultano paralleli agli assi cartesiani.

Esempio 15.2 Sia $\underline{y} = \{y_1, y_2, \ldots, y_n\}$, un c.c.s. da una variabile casuale di **Weibull**, con funzione di densità $p(y; \gamma, \lambda) = \lambda\gamma y^{\gamma-1} e^{-\lambda y^\gamma}$, per $y > 0$, con $\gamma > 0$ e $\lambda > 0$ parametri ignoti. (**Quali sono spazio parametrico e spazio campionario?**) In questo caso, posto $\theta = (\gamma, \lambda)$, per la funzione di verosimiglianza vale l'espressione

$$L(\theta; \underline{y}) = \prod_{i=1}^{n} \lambda\gamma y_i^{\gamma-1} e^{-\lambda y_i^\gamma} \propto \lambda^n \gamma^n \left(\prod_{i=1}^{n} y_i \right)^\gamma e^{-\lambda \sum_{i=1}^n y_i^\gamma}.$$

Quindi,

$$l(\theta; \underline{y}) = n \log \lambda + n \log \gamma + \gamma \sum_{i=1}^{n} \log y_i - \lambda \sum_{i=1}^{n} y_i^\gamma, \quad [10]$$

e per lo score di verosimiglianza si ha

$$l_*(\theta) = \begin{pmatrix} \frac{\partial l(\theta)}{\partial \gamma} \\ \frac{\partial l(\theta)}{\partial \lambda} \end{pmatrix} = \begin{pmatrix} \frac{n}{\gamma} + \sum_{i=1}^n \log y_i - \lambda \sum_{i=1}^n y_i^\gamma \log y_i \\ \frac{n}{\lambda} - \sum_{i=1}^n y_i^\gamma \end{pmatrix}.$$

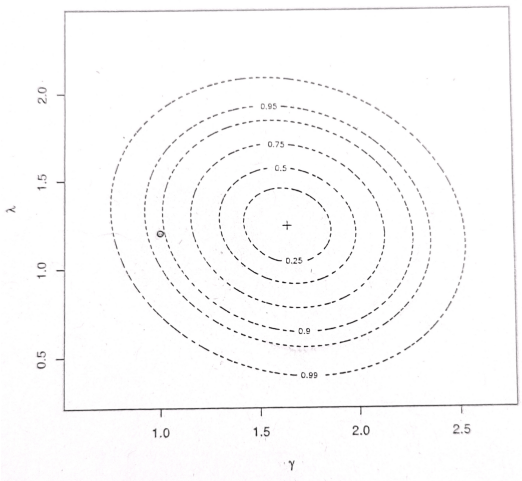
[10] **Si individui una statistica sufficiente per l'inferenza su θ.**

Inoltre la matrice delle derivate seconde vale

$$l_{**}(\theta) = \begin{pmatrix} -\frac{n}{\gamma^2} - \lambda \sum_{i=1}^n y_i^\gamma (\log y_i)^2 & -\sum_{i=1}^n y_i^\gamma \log y_i \\ -\sum_{i=1}^n y_i^\gamma \log y_i & -\frac{n}{\lambda^2}. \end{pmatrix}.$$

Con $n = 20$ e l'osservazione (0,446, 0,604, 2,137, 0,737, 0,996, 1,152, 1,124, 0,137, 0,982, 1,196, 0,841, 0,636, 0,459, 0,947, 0,037, 1,307, 0,858, 0,164, 0,444, 0,623), si ottengono

$$\hat\theta = (\hat\gamma, \hat\lambda) = (1,642, 1,24) \qquad \text{e} \qquad j(\hat\theta) = \begin{pmatrix} 11,76 & 1,63 \\ 1,63 & 13,02 \end{pmatrix}.$$



Il grafico a fianco fornisce le regioni di confidenza per $\theta = (\gamma, \lambda)$ basate su $W_e(\theta; \underline{y}) = 11,76(\gamma - 1,642)^2 + 3,26(\lambda - 1,24)(\gamma - 1,642) + 13,02(\lambda - 1,24)^2$ e l'approssimazione $\chi^2(2)$, per diversi livelli di confidenza nominali (0,25, 0,50, 0,75, 0,90, 0,95, 0,99). Il centro delle ellissi rappresenta la stima di massima verosimiglianza $\hat\theta = (\hat\gamma, \hat\lambda)$ nello spazio parametrico. Come si può notare, in questo caso gli assi delle ellissi che definiscono le regioni risultano non paralleli agli assi cartesiani.

Quando il problema da risolvere è un problema di verifica d'ipotesi su θ, una statistica test può essere ricavata facilmente da $W_e(\theta; \underline{y})$. Infatti, come per il caso $k = 1$, se il problema di verifica d'ipotesi fosse formalizzato dal sistema $H_0 : \theta = \theta_0$ contro $H_1 : \theta \neq \theta_0$, con θ_0 vettore di valori fissati, allora si potrebbe usare la statistica test $W_e(\theta_0; \underline{Y})$, che, sotto H_0, ha distribuzione approssimabile con quella $\chi^2(k)$. Dato che i valori grandi di $W_e(\theta_0; \underline{y})$ sono contrari all'ipotesi nulla, la regione critica del test di livello nominale α è $R_\alpha = (\chi^2_{1-\alpha}(k), +\infty)$.

Con riferimento all'esempio 15.1, supponiamo di considerare l'ipotesi (nulla) $\theta_0 = (\mu_0, \sigma_0^2) = (143, 6, 5^2)$. Risulta $W_e^{oss} = W_e(\theta_0; \underline{y}) = 4,22$ e, poiché il quantile di ordine 0,95 di una variabile $\chi^2(2)$ è 5,99, l'ipotesi nulla non può essere rifiutata (contro l'alternativa più generale), ad un livello nominale del 5%.
Con riferimento all'esempio 15.2, supponiamo di considerare l'ipotesi $\theta_0 = (\gamma_0, \lambda_0) = (1,\ 1,2)$. Nella figura relativa all'esempio, questa ipotesi è rappresentata dal tondino riportato. Risulta $W_e^{oss} = W_e(\theta_0; \underline{y}) = 4,94$, e l'ipotesi nulla non può essere rifiutata, almeno ad un livello nominale del 5% (sarebbe rifiutata ad un livello nominale pari al 10%).

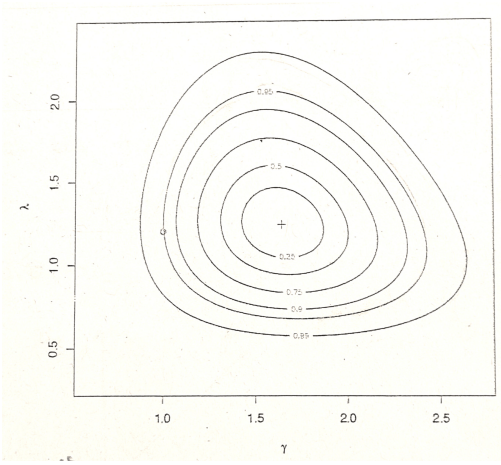
Per completezza, si osservi che nel caso di problema di verifica d'ipotesi, invece di usare $W_e(\theta_0) = (\hat\theta - \theta_0)^\top j(\hat\theta)(\hat\theta - \theta_0)$, si potrebbe usare una qualche "versione alternativa", tipo la statistica $(\hat\theta - \theta_0)^\top i(\theta_0)(\hat\theta - \theta_0)$, per la quale vale sempre l'approssimazione χ^2 sotto H_0.

0.16 Inferenza globale mediante $W(\theta; \underline{y})$

Come nel caso di parametro unidimensionale, anche quando è $k > 1$ le regioni di confidenza per θ costruite usando $W_e(\theta; \underline{y})$ hanno vincoli di simmetria sulla forma (non sempre giustificabili nelle applicazioni) e non raccolgono automaticamente i punti dello spazio parametrico che effettivamente sono a più elevata verosimiglianza attribuita dalle osservazioni. Per ovviare a questi inconvenienti, se si accetta il prezzo di una (di norma) maggiore complessità computazionale, si può optare per l'uso della quantità log-rapporto di verosimiglianza $W(\theta; \underline{Y}) = 2[l(\hat{\theta}; \underline{Y}) - l(\theta; \underline{Y})]$ per ottenere regioni di confidenza. Tale quantità è approssimativamente pivotale sotto θ (e sotto condizioni di regolarità), con distribuzione approssimante ancora $\chi^2(k)$. Le regioni di confidenza per θ che essa produce sono automaticamente regioni interne a Θ e hanno la forma

$$\hat{\Theta}(\underline{y}) = \left\{\theta : W(\theta; \underline{y}) \leq \chi^2_{1-\alpha}(k)\right\},$$

se $1 - \alpha$ è il livello di confidenza nominale richiesto.



Nell'esempio 15.2, dove l'osservazione campionaria è costituita da un c.c.s. da una distribuzione di Weibull con $\theta = (\gamma, \lambda)$ si ha che

$$\begin{aligned} W(\theta; \underline{y}) &= 2\Big[n\log(\hat{\lambda}/\lambda) + n\log(\hat{\gamma}/\gamma) \\ &+ (\hat{\gamma} - \gamma)\sum_{i=1}^{n}\log y_i - \hat{\lambda}\sum_{i=1}^{n} y_i^{\hat{\gamma}} + \lambda\sum_{i=1}^{n} y_i^{\gamma}\Big]. \end{aligned}$$

Per i 20 dati nell'esempio, il grafico a fianco fornisce le regioni di confidenza per θ basate su $W(\theta; \underline{y})$ e l'approssimazione $\chi^2(2)$, a diversi livelli di confidenza nominali (0,25, 0,50, 0,75, 0,90, 0,95, 0,99). Come si può notare, in questo caso le regioni non risultano ellittiche.

Quando il problema da risolvere è un problema di verifica d'ipotesi su θ, con $H_0 : \theta = \theta_0$ contro $H_1 : \theta \neq \theta_0$ e θ_0 vettore di valori fissati, si può ricorrere alla statistica test $W(\theta_0; \underline{Y})$ che, sotto H_0, ha distribuzione approssimabibe con quella $\chi^2(k)$. Anche in questo caso sono i valori grandi della statistica ad essere contrari all'ipotesi nulla; pertanto, la regione critica del test di livello nominale α è $R_\alpha = (\chi^2_{1-\alpha}(k), +\infty)$. Ancora con riferimento all'esempio 15.2 e all'ipotesi $\theta_0 = (\gamma_0, \lambda_0) = (1,\ 1, 2)$, che nella figura sopra è rappresentata sempre dal tondino riportato, risulta $W^{oss} = W(\theta_0; \underline{y}) = 6,068$, per un livello di significatività osservato (approssimato) pari a 0,048.

Esempio 16.1 Un apparecchio elettronico è programmato per rispondere ad un impulso facendo apparire casualmente su uno schermo un ideogramma rappresentante uno dei quattro semi delle carte da poker: cuori, quadri, fiori e picche. Il costruttore afferma che le probabilità con cui la macchina riproduce sullo schermo i quattro semi, ad ogni impulso ricevuto, sono 9/16, 3/16, 3/16 e 1/16, rispettivamente.

Per verificare tale affermazione, in un esperimento si effettuano $n = 556$ prove e si inviano altrettanti impulsi all'apparecchio che riproduce il seme cuore 315 volte, il seme quadri 102 volte, il seme fiori 108 volte, il seme picche 31 volte. L'osservazione sperimentale è quindi costituita dal vettore $\underline{y} = (y_1, y_2, y_3, y_4) = (315, 102, 108, 31)$ e, considerata la natura dell'esperimento, è facile scegliere per la variabile $\underline{Y}$ che genera l'osservazione il modello multinomiale a 4 celle, di indice n e parametro $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$. Ricordiamo che θ rappresenta il vettore delle probabilità con cui si realizzano le 4 "modalità" caratterizzanti le 4 celle. La funzione del modello è, in questo caso,

$$p_{\underline{Y}}(\underline{y}; \theta) = \binom{n}{y_1\ y_2\ y_3\ y_4} \theta_1^{y_1} \theta_2^{y_2} \theta_3^{y_3} \theta_2^{y_4},$$

e, dunque, $L(\theta; \underline{y}) \propto \theta_1^{y_1} \theta_2^{y_2} \theta_3^{y_3} \theta_2^{y_4}$. Teniamo presente che $y_1+y_2+y_3+y_4 = n$ e $\theta_1+\theta_2+\theta_3+\theta_4 = 1$, con $\theta_i \in (0, 1)$, $i = 1, 2, 3, 4$. **(Quali sono lo spazio parametrico e lo spazio campionario?)** Pertanto, possiamo dedurre che $\dim(\Theta) = 3$, ovvero che la dimensione "effettiva" di θ è $k = 3$. Per la log-verosimiglianza vale l'espressione

$$\begin{aligned} l(\theta; \underline{y}) &= y_1 \log \theta_1 + y_2 \log \theta_2 + y_3 \log \theta_3 + y_4 \log \theta_4 \\ &= y_1 \log \theta_1 + y_2 \log \theta_2 + y_3 \log \theta_3 + y_4 \log(1 - \theta_1 - \theta_2 - \theta_3) \end{aligned}$$

e lo score è un vettore di dimensione 3, il cui i-esimo elemento è

$$\frac{y_i}{\theta_i} - \frac{y_4}{\theta_4},$$

$i = 1, 2, 3$. Sommando i membri delle tre equazioni $\frac{y_i}{\theta_i} = \frac{y_4}{\theta_4}$ che compongono il sistema di equazioni di verosimiglianza e tenedo conto che $y_1+y_2+y_3 = n-y_4$ e che $\theta_1+\theta_2+\theta_3 = 1-\theta_4$, otteniamo un'equazione in θ_4,

$$(n - y_4)\theta_4 = y_4(1 - \theta_4),$$

che risolta porta a $\hat{\theta}_4 = y_4/n$. Di conseguenza, tornando alle tre equazioni del sistema, si ottengono $\hat{\theta}_i = y_i/n$, per $i = 1, 2, 3$. Quindi la stima di massima verosimiglianza vale $\hat{\theta} = (y_1/n, y_2/n, y_3/n, y_4/n)$.
Stabilire l'attendibilità dell'affermazione del costruttore significa, in questo caso, verificare l'ipotesi $H_0 : \theta = \theta_0$, con $\theta_0 = (\theta_{1_0}, \theta_{2_0}, \theta_{3_0}, \theta_{4_0}) = (9/16, 3/16, 3/16, 1/16)$. È facile verificare che

$$l(\hat{\theta}; \underline{y}) - l(\theta_0; \underline{y}) = \sum_{i=1}^{4} y_i \log \frac{\hat{\theta}_i}{\theta_{i_0}}$$

e, con i dati risulta $W^{oss} = W(\theta_0; \underline{y}) = 0,62$. Il livello di significatività osservato approssimato corrispondente, cioè la $\Pr\{V \geq 0,62\}$ con $V \sim \chi^2(3)$, è superiore a 0,8; quindi il supporto dei dati all'affermazione del costruttore appare molto forte.

0.17 Inferenza parziale basata su quantità di tipo Wald: interesse per un singolo elemento di θ

Sia ancora $\underline{y} = \{y_1, y_2, \ldots, y_n\}$ l'osservazione campionaria e sia $\mathcal{F}$ il modello scelto per la variabile $\underline{Y}$, di cui $\underline{y}$ è realizzazione. Sia $\theta \in \Theta$ il parametro che indicizza il modello, con $k = \dim(\Theta) > 1$.

Indichiamo con $\theta_1, \theta_2, \ldots, \theta_k$ gli elementi del vettore θ, e supponiamo di essere interessati ad un singolo elemento, per esempio θ_1. Come possiamo costruire intervalli di confidenza per θ_1, o risolvere problemi di verifica d'ipotesi che riguardano tale entità, tenendo conto della presenza di $k-1$ elementi di disturbo?

Se siamo sotto condizioni di regolarità, una soluzione semplice discende dai risultati asintotici (consistenza e normalità) che valgono per lo stimatore di massima verosimiglianza $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$. Infatti, se indichiamo con $[j^{-1}(\hat{\theta})]_{11}$ l'elemento di posizione (1,1) nella matrice $j^{-1}(\hat{\theta})$, il risultato $\hat{\theta}(\underline{Y}) \dot{\sim} N_k(\theta, j^{-1}(\hat{\theta}))$, sotto θ, implica $\hat{\theta}_1(\underline{Y}) \dot{\sim} N(\theta_1, [j^{-1}(\hat{\theta})]_{11})$, sotto θ_1. Ciò permette di considerare la quantità

$$r_{eP}(\theta_1; \underline{Y}) = \frac{\hat{\theta}_1 - \theta_1}{\sqrt{[j^{-1}(\hat{\theta})]_{11}}},$$

che è approssimativamente pivotale per l'inferenza su θ_1, con distribuzione approssimante normale standard (sotto θ_1). Si tratta di una **quantità di tipo Wald parziale**. Essa va usata alla stregua di r_e nel caso unidimensionale e permette di ottenere intervalli di confidenza per il parametro di interesse θ_1. Tali intervalli hanno la solita forma: sono centrati sulla stima $\hat{\theta}_1$, con estremi $\hat{\theta}_1 \pm z_{1-\alpha/2}\sqrt{[j^{-1}(\hat{\theta})]_{11}}$ se $1-\alpha$ è il livello di confidenza nominale richiesto.
Come nel caso unidimensionale ($k = 1$), anche in quello che stiamo trattando si può considerare la quantità $W_{eP}(\theta_1; \underline{Y}) = [r_{eP}(\theta_1; \underline{Y})]^2$ che è approssimativamente pivotale per l'inferenza su θ_1, con distribuzione approssimante $\chi^2(1)$ sotto θ_1. Per ogni livello nominale fissato, gli intervalli

$$\hat{\Theta}_P(\underline{y}) = \left\{\theta_1 \in \Theta_1 : W_{eP}(\theta_1; \underline{y}) \leq \chi^2_{1-\alpha}(1)\right\},$$

dove Θ_1 indica il sottospazio dello spazio parametrico in cui varia θ_1, coincidono con quelli costruiti tramite $r_{eP}(\theta_1; \underline{y})$.

Esempio 17.1 Riprendiamo il caso di c.c.s. da una normale. Sia $\underline{y} = \{y_1, y_2, \ldots, y_n\}$ un c.c.s. da una distribuzione $N(\mu, \sigma^2)$. Sia $\theta = (\mu, \sigma^2)$ il parametro che indicizza il modello per $\underline{Y}$, e sia $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$, con $\hat{\mu}$ e $\hat{\sigma}^2$ rispettivamente media e varianza campionarie. Possiamo ricavare

$$j^{-1}(\hat{\theta}) = \begin{pmatrix} \frac{\hat{\sigma}^2}{n} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{n} \end{pmatrix}.$$

Allora, se l'interesse fosse rivolto alla media μ, potremmo considerare la quantità

$$r_{eP}(\mu; \underline{Y}) = \frac{\hat{\mu} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}},$$

che è approssimativamente pivotale sotto μ, con distribuzione approssimante $N(0,1)$. Da essa si potrebbero ottenere intervalli di confidenza approssimati per μ. In realtà, sappiamo che abbiamo una quantità esattamente pivotale per l'inferenza sulla media μ in questo contesto. Tale quantità è

$$\frac{\hat{\mu} - \mu}{\sqrt{\frac{S^2}{n}}},$$

con distribuzione esatta $t(n-1)$ sotto μ. È quindi ragionevole utilizzare quest'ultima quantità per costruire intervalli di confidenza per la media di una popolazione normale (con varianza ignota).

Se l'interesse fosse rivolto alla varianza σ^2, potremmo utilizzare la quantità

$$r_{eP}(\sigma^2; \underline{Y}) = \frac{\hat{\sigma}^2 - \sigma^2}{\sqrt{\frac{2\hat{\sigma}^4}{n}}},$$

che è approssimativamente pivotale sotto σ^2, con distribuzione approssimante $N(0,1)$, e che produrrebbe intervalli di confidenza del tipo $\left(\hat{\sigma}^2 - z_{1-\alpha/2}\sqrt{\frac{2\hat{\sigma}^4}{n}},\ \hat{\sigma}^2 + z_{1-\alpha/2}\sqrt{\frac{2\hat{\sigma}^4}{n}}\right)$. Ad ogni modo, anche in questo caso conosciamo una quantità esattamente pivotale per l'inferenza su σ^2, che è $n\hat{\sigma}^2/\sigma^2$, con distribuzione esatta $\chi^2(n-1)$ sotto σ^2.

Esempio 17.2 Riprendiamo la situazione delineata nell'esempio 15.2, con l'osservazione $\underline{y} = \{y_1, y_2, \ldots, y_n\}$ costituita da un c.c.s. da una variabile casuale di Weibull, con parametro bidimensionale $\theta = (\gamma, \lambda)$, con $\gamma > 0$ e $\lambda > 0$. Supponiamo di essere interessati alla costruzione di un intervallo di cofidenza per γ.

Dato che

$$\hat{\theta} = (\hat{\gamma}, \hat{\lambda}) = (1,642, 1,24) \qquad \text{e} \qquad j(\hat{\theta}) = \begin{pmatrix} 11,76 & 1,63 \\ 1,63 & 13,02 \end{pmatrix},$$

possiamo ottenere

$$j^{-1}(\hat{\theta}) = \frac{1}{150,45}\begin{pmatrix} 13,02 & -1,63 \\ -1,63 & 11,76 \end{pmatrix}.$$

Ne segue che, utilizzando $r_{eP}(\gamma; \underline{y})$, un intervallo di confidenza per γ di livello nominale 0,95 ha estremi $1,642 \pm 1,96\sqrt{0,0865}$.

Quando il problema da affrontare è un problema di verifica d'ipotesi su θ_1, con ipotesi nulla H_0 : $\theta_1 = \theta_{1_0}$ e θ_{1_0} valore fissato, si possono derivare facilmente le statistiche test

$$r_{eP}(\theta_{1_0}; \underline{Y}) = \frac{\hat{\theta}_1 - \theta_{1_0}}{\sqrt{[j^{-1}(\hat{\theta})]_{11}}}$$

e $W_{eP}(\theta_{1_0}; \underline{Y}) = [r_{eP}(\theta_{1_0}; \underline{Y})]^2$, che sotto H_0 hanno distribuzione approssimabile con quella $N(0,1)$ e quella $\chi^2(1)$, rispettivamente. Tali statistiche test vanno usate come le corrispondenti statistiche viste nel caso unidimensionale ($k = 1$). Quindi, in particolare, la forma della regione di rifiuto dipende dal tipo di alternativa contrapposta ad H_0, e $W_{eP}(\theta_{1_0}; \underline{y})$ può essere usata solo con alternative bilaterali.

0.18 Inferenza parziale basata sul log-rapporto di verosimiglianza: interesse per un singolo elemento di θ

Soluzioni per l'inferenza sul singolo elemento, diciamo θ_1, di θ, alternative a quelle viste nella sezione precedente, possono essere trovate nell'ambito delle quantità di tipo log-rapporto di verosimiglianza. In particolare, nel caso unidimensionale, cioè con $k = 1$, abbiamo definito la quantità $W(\theta; \underline{y}) = 2[l(\hat{\theta}; \underline{y}) - l(\theta; \underline{y})]$, che misura la caduta della log-verosimiglianza quando si passa dal suo punto di massimo al punto θ dello spazio parametrico, che (data l'osservazione campionaria) è l'argomento stesso della funzione $W(\theta; \underline{y})$. Come si può estendere tale quantità al caso in cui il parametro di interesse θ_1 è ancora unidimensionale ma si è in presenza di $k-1$ elementi di disturbo?

A ben vedere, almeno in linea di principio, la risposta a questo quesito dovrebbe apparire semplice. Infatti la questione diventa, sostanzialmente, stabilire quale valore di log-verosimiglianza attribuire a quello che ora diventa l'argomento della funzione (data l'osservazione), ossia θ_1: si può decidere di attribuire il valore massimo che la log-verosimiglianza raggiunge (con θ_1 fissato) al variare dei restanti $k-1$ elementi di θ in Θ. In definitiva, possiamo definire la quantità

$$W_P(\theta_1; \underline{y}) = 2\left[l(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k; \underline{y}) - \max_{\theta_2,\ldots,\theta_k} l(\theta_1, \theta_2, \ldots, \theta_k; \underline{y})\right],$$

dove, per facilitare la presentazione, ora gli elementi di θ appaiono distinti nella espressione di $l(\cdot;\cdot)$. Se indichiamo con $\hat{\theta}_{h\theta_1}$, con $h = 2, 3, \ldots, k$ la stima di massima verosimiglianza dell'elemento $h-$esimo di θ, vincolata al valore fissato di θ_1, allora possiamo esprimere $W_P(\theta_1; \underline{y})$ nel modo seguente

$$W_P(\theta_1; \underline{y}) = 2\left[l(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k; \underline{y}) - l(\theta_1, \hat{\theta}_{2\theta_1}, \ldots, \hat{\theta}_{k\theta_1}; \underline{y})\right].$$

Dovrebbe essere chiaro che questa estensione mantiene inalterata la logica caratterizzante la quantità "originale" (quella del caso $k = 1$) e che, quindi, almeno in linea di principio, la quantità $W_P(\theta_1; \underline{y})$ potrebbe essere usata in un modo che già conosciamo per fare inferenza su θ_1. Perché in linea di principio? Perché abbiamo la necessità di conoscere la distribuzione di $W_P(\theta_1; \underline{y})$, almeno approssimativamente, sotto θ_1. Fortunatamente, si dimostra che, **sotto condizioni di regolarità** e sotto θ_1

$$W_P(\theta_1; \underline{Y}) \dot{\sim} \chi^2(1).$$

Questo risultato permette di affermare che $W_P(\theta_1; \underline{Y})$ è quantità approssimativamente pivotale per l'inferenza su θ_1, e di ottenere intervalli (più generalmente regioni) di confidenza per il parametro di interesse θ_1 come

$$\hat{\Theta}_P(\underline{y}) = \left\{\theta_1 : W_P(\theta_1; \underline{y}) \leq \chi^2_{1-\alpha}(1)\right\},$$

se $1-\alpha$ è il livello di confidenza nominale richiesto. Vale la pena osservare che, data l'osservazione campionaria, l'argomento θ_1 entra nella funzione $W_P(\theta_1; \underline{y})$ non solo direttamente ma anche attraverso i valori delle stime vincolate degli elementi di disturbo. A livello computazionale, tipicamente ciò comporta una ulteriore complicazione per la costruzione degli intervalli di confidenza basati su $W_P(\theta_1; \underline{y})$, costruzione che può diventare parecchio "onerosa" soprattutto se confrontata con quella

semplice che caratterizza gli intervalli ottenibili da $r_{eP}(\theta_{1_0}; \underline{y})$.

Esempio 18.1 Sia $\underline{y} = \{y_1, y_2, \ldots, y_n\}$ un c.c.s. da una variabile casuale $N(\mu, \sigma^2)$. Abbiamo visto che, posto $\theta = (\mu, \sigma^2)$, per la log-verosimiglianza vale l'espressione

$$l(\theta; \underline{y}) = l(\mu, \sigma^2; \underline{y}) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

Ora, se nell'espressione di $l(\mu, \sigma^2; \underline{y})$ fissiamo μ, otteniamo una funzione solo di σ^2 (data l'osservazione $\underline{y}$), la cui derivata rispetto a σ^2 risulta essere

$$-\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^4}.$$

La radice dell'equazione

$$-\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^4} = 0$$

si ricava facilmente ed è $\hat{\sigma}^2_\mu = \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{n}$; essa rappresenta la **stima di massima verosimiglianza vincolata** di σ^2, vincolata cioè al valore fissato μ per la media. Di conseguenza abbiamo

$$\begin{aligned} W_P(\mu; \underline{y}) &= 2[l(\hat{\mu}, \hat{\sigma}^2; \underline{y}) - l(\mu, \hat{\sigma}^2_\mu; \underline{y})] \\ &= -n\log\hat{\sigma}^2 - (1/\hat{\sigma}^2)\sum_{i=1}^{n}(y_i - \hat{\mu})^2 + n\log\hat{\sigma}^2_\mu + (1/\hat{\sigma}^2_\mu)\sum_{i=1}^{n}(y_i - \mu)^2 \\ &= n\log\frac{\hat{\sigma}^2_\mu}{\hat{\sigma}^2} = n\log\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = n\log\left[1 + \frac{(\bar{y}-\mu)^2}{\hat{\sigma}^2}\right]. \end{aligned}$$

Quando il problema da affrontare è un problema di verifica d'ipotesi su θ_1, con ipotesi nulla H_0 : $\theta_1 = \theta_{1_0}$ (θ_{1_0} valore fissato) e ipotesi alternativa bilaterale, si può ricorrere alla statistica test log-rapporto di verosimiglianza parziale

$$W_P(\theta_{1_0}; \underline{Y}) = 2\left[l(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k; \underline{Y}) - l(\theta_{1_0}, \hat{\theta}_{2\theta_{1_0}}, \ldots, \hat{\theta}_{k\theta_{1_0}}; \underline{Y})\right],$$

che, sotto H_0 (e sotto condizioni di regolarità), ha distribuzione approssimabile con la distribuzione $\chi^2(1)$. Essendo i valori grandi della statistica ad essere contrari all'ipotesi nulla, si rifiuta H_0 ad un livello nominale α se il valore osservato $W_P^{oss} = W_P(\theta_{1_0}; \underline{y})$ è più grande del quantile $\chi^2_{1-\alpha}(1)$.
Per affrontare problemi di verifica d'ipotesi con alternative unilaterali si può definire la versione **radice con segno parziale**

$$r_P(\theta_{1_0}; \underline{Y}) = sgn(\hat{\theta}_1 - \theta_{1_0})\sqrt{W_P(\theta_{1_0}; \underline{Y})},$$

che sotto H_0 ha distribuzione approssimante $N(0, 1)$.

Con riferimento all'esempio 18.1, assumiamo sia $n = 9$ e di disporre dell'informazione di sintesi: $\bar{y} = 4$ e $\hat{\sigma}^2 = (1/9)\sum_{i=1}^{9}(y_i - \bar{y})^2 = 1,5^2$. Consideriamo il problema di verifica d'ipotesi

con $H_0 : \mu = \mu_0 = 3$ contro $H_1 : \mu > \mu_0$. Si può allora calcolare $W_P^{oss} = W_P(\mu_0; \underline{y}) = 9\log(1 + (1/1,5)^2) = 3,3$, da cui si ricava $r_P^{oss} = r_P(\mu_0; \underline{y}) = +\sqrt{3,3} = 1,8136$. Questo valore porta a rifiutare H_0, ad un livello nominale del 5%.

0.19 Inferenza parziale: interesse per un insieme di elementi di θ

Supponiamo ora di essere interessati a un qualche sottoinsieme di elementi (più di uno) del vettore θ che indicizza il modello statistico scelto per la variabile $\underline{Y}$, che descrive l'osservazione $\underline{y}$ di cui disponiamo. Senza perdita di generalità, possiamo partizionare $\theta \in \Theta$ come $\theta = (\tau, \eta)$, con τ entità di interesse e η entità di disturbo. Se $k = \dim(\Theta)$, indichiamo con h $(1 < h < k)$ la dimensione di τ.

Per fare inferenza su τ possiamo estendere le quantità viste nel caso specifico, con $h = 1$, trattato finora. In effetti, possiamo pensare alle due quantità

$$W_{eP}(\tau; \underline{Y}) = (\tau - \hat{\tau})^\top j_{\tau\tau}(\hat{\theta})(\tau - \hat{\tau}) \quad \text{e} \quad W_P(\tau; \underline{Y}) = 2[l(\hat{\tau}, \hat{\eta}) - l(\tau, \hat{\eta}_\tau)],$$

per le quali, sotto τ e sotto le condizioni di regolarità vale l'approssimazione χ^2. Nell'espressione di $W_{eP}(\tau; \underline{y})$, $j_{\tau\tau}(\hat{\theta})$ indica la stima della parte della matrice di informazione osservata $j(\theta)$ relativa alla componente τ di θ ($j_{\tau\tau}(\hat{\theta})$ è quindi una sotto-matrice di $j(\hat{\theta})$). Nell'espressione di $W_P(\tau; \underline{y})$, $\hat{\eta}_\tau$ indica la stima di massima verosimiglianza del parametro di disturbo η, vincolata al valore fissato τ per l'entità di interesse.

Le due quantità considerate sono approssimativamente pivotali per l'inferenza su τ, con distribuzione approssimante $\chi^2(h)$ sotto τ. Pertanto, regioni di confidenza di livello nominale $1 - \alpha$ per τ sono fornite da

$$\hat{\Theta}_P(\underline{y}) = \left\{\tau \in \Theta_T : W_{eP}(\tau; \underline{y}) \leq \chi^2_{1-\alpha}(h)\right\} \quad \text{e} \quad \hat{\Theta}_P(\underline{y}) = \left\{\tau : W_P(\tau; \underline{y}) \leq \chi^2_{1-\alpha}(h)\right\},$$

dove con Θ_T si è indicato il sottospazio di Θ in cui varia τ. Le regioni ottenute da $W_{eP}(\tau; \underline{y})$ sono centrate sulla stima di massima verosimiglianza $\hat{\tau}$ e sono di forma ellittica se $h = 2$ (più in generale sono ellissoidi). Le regioni ottenute da $W_P(\tau; \underline{y})$ non hanno invece vincoli sulla forma e sono sempre (automaticamente) interne a Θ_T. Sono però tipicamente più "onerose" dal punto di vista computazionale.

Quando il problema da risolvere è un problema di verifica d'ipotesi, con $H_0 : \tau = \tau_0$ contro $H_1 : \tau \neq \tau_0$ e τ_0 vettore di valori fissati, si possono utilizzare le statistiche test $W_{eP}(\tau_0; \underline{Y})$ e $W_P(\tau_0; \underline{Y})$, che sotto H_0 (e sotto le condizioni di regolarità) hanno distribuzione approssimante $\chi^2(h)$. Essendo i valori grandi di $W_{eP}(\tau_0; \underline{y})$ e di $W_P(\tau_0; \underline{y})$ ad essere contrari ad H_0, la regione di rifiuto di livello di significatività nominale α è, per enrambi i test, $R_\alpha = (\chi^2_{1-\alpha}(h), +\infty)$.

0.20 Inferenza parziale: problema di verifica d'ipotesi con vincoli sugli elementi di θ

Il problema che affrontiamo in questa sezione è un problema di verifica d'ipotesi, in cui, in termini molto generali, l'ipotesi nulla fissa una serie di vincoli sul parametro θ (di dimensione $k > 1$) che indicizza il modello statistico parametrico scelto per la variabile $\underline{Y}$, di cui si pensa realizzazione l'intera osservazione campionaria $\underline{y}$.

Immaginiamo di dover valutare l'aderenza delle ossevazioni ad una qualche ipotesi di interesse, H_0, e suppuniamo che tale ipotesi, nella sua articolazione, fissi, di fatto, un certo insieme di vincoli sul parametro θ. Sotto H_0, quindi, θ non sarà libero di "muoversi" su tutto lo spazio parametrico Θ, ma potrà variare in un suo sottospazio, diciamo Θ_0. L'ipotesi nulla (quale essa sia) sarà dunque caratterizzata dal sottospazio Θ_0 ad essa associato, e potrà essere formulata come $H_0 : \theta \in \Theta_0$. Se ad essa contrapponiamo l'alternativa più generale, $H_1 : \theta \notin \Theta_0$, possiamo cercare la soluzione del problema di verifica d'ipotesi nell'ambito di quelle soluzioni che fanno capo al log-rapporto di verosimiglianza.

In effetti, possiamo definire la statistica test

$$W_P^{H_0} = 2[l(\hat{\theta}) - l(\hat{\theta}_0)],$$

dove con $\hat{\theta}_0$ indichiamo la stima di massima verosimiglianza di θ sotto H_0, cioè quando il parametro varia solo in Θ_0. Data l'osservazione $\underline{y}$, tale statistica misura ancora una caduta di log-verosimiglianza, dal massimo assoluto al "massimo che i dati attribuiscono ad H_0". Se tale caduta fosse eccessiva, i dati indicherebbero contrarietà ad H_0. Come al solito, per valutare la "significatività" della caduta di log-verosimiglianza osservata, c'è bisogno di conoscere, almeno in termini approssimati, la distribuzione di $W_P^{H_0}$ (inteso come oggetto casuale) sotto H_0. Quando sono valide le condizioni di regolarità, si dimostra che, sotto H_0,

$$W_P^{H_0}(\underline{Y}) \dot{\sim} \chi^2(d),$$

con $d = \dim(\Theta) - \dim(\Theta_0)$. Pertanto la regione di rifiuto del test basato su $W_P^{H_0}$, di livello di significatività nominale α, è $R_\alpha = (\chi^2_{1-\alpha}(d), +\infty)$.

Esempio 20.1 In un esperimento per uno studio in agraria, si è interessati a valutare se il tasso di germinazione di una certa specie vegetale sia lo stesso in quattro terreni, di pari estensione, situati in zone diverse di una certa regione. Su ogni terreno vengono sparsi 1000 semi di quella specie e dopo un certo periodo si contano i nuovi germogli. L'osservazione campionaria è costituita dal vettore $\underline{y} = (200, 178, 192, 213)$.
Data la natura dell'esperimento, è ragionevole ritenere che il vettore casuale $\underline{Y}$ che genera l'osservazione sia a componenti indipendenti e che ciascuna componente sia binomiale di indice 1000, cioè, $Y_i \sim Bi(n, \theta_i)$ con $n = 1000$ e $\theta_i \in (0, 1)$, $i = 1, 2, 3, 4$. In questo modo rimane fissato un modello statistico $\mathcal{F}$ per $\underline{Y}$, con parametro $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ e funzione di verosimiglianza

$$L(\theta; \underline{y}) \propto \prod_{i=1}^{4} \theta_i^{y_i}(1 - \theta_i)^{n-y_i}.$$

Il quesito di interesse può essere formalizzato con l'ipotesi $H_0 : \theta_1 = \theta_2 = \theta_3 = \theta_4$ e, siccome sotto H_0 rimane ignoto solo il valore comune ai quattro elementi di θ, è chiaro che $\dim(\Theta_0) = 1$. Quindi, in questo caso, sono soddisfatte le condizioni di regolarità e la statistica test $W_P^{H_0}(\underline{Y})$ ha distribuzione, sotto H_0, approssimabile con la legge $\chi^2(3)$, essendo $\dim(\Theta) - \dim(\Theta_0) = 3$.
Per calcolare il valore della statistica occorre ottenere la stima di massima verosimiglianza non vincolata $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4)$ e la stima del valore comune, diciamo θ_*, per gli elementi di θ sotto H_0. Dalla espressione di $L(\theta; \underline{y})$, si ha

$$l(\theta; \underline{y}) = l(\theta_1, \theta_2, \theta_3, \theta_4; \underline{y}) = \sum_{i=1}^{4} [y_i \log \theta_i + (n - y_i) \log(1 - \theta_i)],$$

e lo score di verosimiglianza risulta essere un vettore di dimensione quattro, il cui generico i-esimo elemento ha espressione

$$\frac{y_i}{\theta_i} - \frac{n - y_i}{1 - \theta_i}.$$

Come si può notare, ogni elemento dello score coinvolge un solo elemento di θ. La soluzione del sistema di equazioni di verosimiglianza è quindi immediata e porta ad ottenere le stime (non vincolate) $\hat{\theta}_i = y_i/n$, per $i = 1, 2, 3, 4$.
Sotto H_0, cioè quando è ignoto solo il valore comune θ_*, la log-verosimiglianza diventa

$$l(\theta_*, \theta_*, \theta_*, \theta_*; \underline{y}) = \log \theta_* \sum_{i=1}^{4} y_i + \log(1 - \theta_*) \left(4n - \sum_{i=1}^{4} y_i \right).$$

Derivando rispetto a θ_*, ponendo uguale a zero e risolvendo l'equazione si ottiene la stima $\hat{\theta}_* = \frac{\sum_{i=1}^{4} y_i}{4n}$. In definitiva, con i dati abbiamo $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4) = (0, 2, 0, 178, 0, 192, 0, 213)$ e $\hat{\theta}_0 = (\hat{\theta}_*, \hat{\theta}_*, \hat{\theta}_*, \hat{\theta}_*) = (0, 196, 0, 196, 0, 196, 0, 196)$. Inoltre, $l(\hat{\theta}; \underline{y}) = -1975, 76$, $l(\hat{\theta}_0; \underline{y}) = -1977, 81$ e $W_P^{H_0} = 4, 1$. A tale valore corrisponde un livello di significatività osservato (approssimato) maggiore di 0,2. Concludiamo che i dati non portano a rifiutare l'ipotesi nulla (equivalenza dei quattro terreni in termini di "fertilità").

0.21 Alcune considerazioni finali

La maggior parte delle tecniche viste in queste pagine si basano sui risultati asintotici che caratterizzano le quantità di verosimiglianza e che valgono sotto le condizioni di regolarità. Le procedure inferenziali derivanti da tali tecniche godono di una buona generalità, ma sono procedure approssimate, con un "errore" di approssimazione non controllabile, che può risultare rilevante in campioni di piccole dimensioni (soprattutto se la dimensione del parametro non è piccola).
Comunque, come abbiamo già avuto modo di osservare in precedenza, è importante tenere presente che quando si usa una quantità derivante dalla funzione di verosimiglianza per fare inferenza, che si tratti di quantità pivotale o statistica test, si può sempre provare ad evitare la "scorciatoia" rappresentata dal risultato asintotico e cercare di studiarne il comportamento esatto (la distribuzione esatta).

Per esempio, se il problema da risolvere fosse un problema di verifica d'ipotesi, volendo fare inferenza non approssimata, si potrebbe provare ad individuare la distribuzione esatta della statistica test sotto l'ipotesi nulla. In talune situazioni, invece, studiando l'espressione della statistica test, si potrebbe stabilire un legame funzionale con un'altra statistica, per la quale si riesce ad individuare la distribuzione sotto H_0. A titolo illustrativo, se riprendiamo il problema discusso alla fine della sezione 18, relativo all'esempio 18.1, per la statistica test $W_P(\mu_0; \underline{y})$ possiamo scrivere

$$\begin{aligned} W_P(\mu_0; \underline{y}) &= n \log \left[1 + \frac{(\bar{y} - \mu_0)^2}{\hat{\sigma}^2} \right] \\ &= n \log \left[1 + \frac{t^2}{n-1} \right], \end{aligned}$$

dove

$$t^2 = \frac{(n-1)(\bar{y} - \mu_0)^2/\sigma^2}{\hat{\sigma}^2/\sigma^2}.$$

Ne risulta che $W_P(\mu_0; \underline{y})$ è funzione monotona crescente di t^2. Quindi, rifiutare $H_0 : \mu = \mu_0$ (contro $H_1 : \mu \neq \mu_0$) per valori grandi della statistica test log-rapporto di verosimiglianza (parziale), equivale a rifiutare H_0 per valori grandi di t^2. Inoltre, la variabile T^2, di cui è realizzazione t^2, ha distribuzione esatta, sotto H_0, $F(1, n-1)$ **(perché?)** e questo permette di effettuare un test esatto: si rifiuta H_0 al livello esatto α se $t^2_{oss} > F_{1-\alpha}(1, n-1)$, dove $F_{1-\alpha}(1, n-1)$ indica il quandile di ordine $1-\alpha$ della distribuzione $F(1, n-1)$.

Un'ultima considerazione riguarda l'uso della statistica $W_P^{H_0}$, relativa al modello multinomiale, come test di bontà di adattamento.
Supponiamo di disporre dell'osservazione campionaria (x_i, y_i), $i = 1, 2, \ldots, q$, dove, in ciascuna coppia, x_i rappresenta il valore osservato del carattere considerato, e y_i la frequenza con la quale (cioè quante volte) quel valore si osserva. Siano, per esempio, $q = 5$, $\underline{x} = (x_1, x_2, x_3, x_4, x_5) = (0, 1, 2, 3, 4)$ e $\underline{y} = (109,\ 65,\ 22,\ 3,\ 1)$: si vuole stabilire se sia ragionevole ipotizzare il modello di Poisson come modello generatore dei 200 dati.
In alternativa al classico test di Pearson, si può in questo caso ricorrere alla statistica $W_P^{H_0}$, relativa al modello multinomiale, come test di bontà di adattamento. Infatti se consideriamo $\underline{y} = (109,\ 65,\ 22,\ 3,\ 1)$ come osservazione campionaria e pensiamo alla variabile $\underline{Y}$ come variabile multinomiale a 5 celle, indice $n = 200$ e vettore dei parametri $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$, il quesito d'interesse può essere formalizzato con l'ipotesi che fissa $\theta_1 = \theta_1(\lambda) = \Pr\{X = 0\} = e^{-\lambda}$, $\theta_2 = \theta_2(\lambda) = \Pr\{X = 1\} = \lambda e^{-\lambda}$, $\theta_3 = \theta_3(\lambda) = \Pr\{X = 2\} = (1/2)\lambda^2 e^{-\lambda}$, $\theta_4 = \theta_4(\lambda) = \Pr\{X = 3\} = (1/3!)\lambda^3 e^{-\lambda}$ e $\theta_5 = \theta_5(\lambda) = \Pr\{X = 4\} = (1/4!)\lambda^4 e^{-\lambda}$, dove $\lambda > 0$ indica il parametro (ignoto) che caratterizza una variabile casuale di Poisson[11]. Allora, la statistica test

$$W_P^{H_0} = 2[l(\hat{\theta}; \underline{Y}) - l(\hat{\theta}_0; \underline{Y})],$$

con $l(\cdot; \cdot)$ log-verosimiglianza per il modello multinomiale e $\hat{\theta}_0 = \theta(\hat{\lambda})$ stima di massima verosimiglianza vincolata all'ipotesi $H_0 : X \sim \text{Poisson}$, ha distribuzione approssimante $\chi^2(d)$, sotto H_0, con $d = \dim(\Theta) - \dim(\Theta_0) = 4 - 1 = 3$.

[11] Si osservi che aver definito $\theta_5(\lambda) = \Pr\{X = 4\}$ e non $\theta_5(\lambda) = \Pr\{X \geq 4\}$ è elemento trascurabile in pratica.

Esercizi riepilogativi

1. Sia $x_1, x_2, \ldots, x_n$ un campione casuale semplice da una variabile X con funzione di densità

$$p_X(x; \mu, \lambda) = \left[\frac{\lambda}{2\pi x^3}\right]^{1/2} \exp\left\{\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right\}$$

per $x > 0$, con $\mu > 0$ e $\lambda > 0$ parametri ignoti.

(a) Si ottenga la funzione di log-verosimiglianza, $l(\theta)$, individuando una statistica sufficiente per l'inferenza su $\theta = (\mu, \lambda)$.

(b) Si ottenga lo stimatore di massima verosimiglianza $\hat{\theta} = (\hat{\mu}, \hat{\lambda})$. Si mostri che $\hat{\mu}$ è non distorto.

(c) Si fornisca un'approssimazione per la distribuzione di $\hat{\theta}$.

(d) Con $n = 10$ e i dati

$$3,85 \;\; 0,51 \;\; 0,61 \;\; 1,43 \;\; 0,36 \;\; 7,48 \;\; 6,23 \;\; 1,29 \;\; 0,37 \;\; 2,75,$$

si verifichi (ad un livello approssimato del 5%) l'ipotesi $H_0 : \mu = 3$, utilizzando la statistica W_P.

(e) Si mostri che $Var(X) = \mu^3/\lambda$. Come si potrebbe costruire un intervallo di confidenza per $Var(X)$ utilizzando una quantità basata sulla funzione di verosimiglianza?

Soluzione

(a) Per quanto riguarda la funzione di verosimiglianza, si ha

$$L(\mu, \lambda) \propto \lambda^{n/2} \exp\left[-\frac{\lambda}{2\mu^2}\sum_i \frac{(x_i-\mu)^2}{x_i}\right].$$

Quindi, passando al logaritmo,

$$\begin{aligned} l(\mu, \lambda) &= \frac{n}{2}\log\lambda - \frac{\lambda}{2\mu^2}\sum_i \frac{(x_i-\mu)^2}{x_i} \\ &= \frac{n}{2}\log\lambda - \frac{\lambda}{2\mu^2}\sum_i x_i + \frac{n\lambda}{\mu} - \frac{\lambda}{2}\sum_i \frac{1}{x_i} \end{aligned}$$

e una statistica sufficiente per l'nferenza su $\theta = (\mu, \lambda)$ è $\left(\sum_i x_i, \sum_i \frac{1}{x_i}\right)$.

(b) Derivando la funzione di log-verosimiglianza, si ottiene la funzione punteggio $l_*(\mu, \lambda) = \left(\frac{\partial l(\mu,\lambda)}{\partial \mu}, \frac{\partial l(\mu,\lambda)}{\partial \lambda}\right)^\top$, con

$$\begin{aligned} \frac{\partial l(\mu, \lambda)}{\partial \mu} &= \frac{n\lambda\bar{x}}{\mu^3} - \frac{n\lambda}{\mu^2} \\ \frac{\partial l(\mu, \lambda)}{\partial \lambda} &= \frac{n}{2\lambda} - \frac{1}{2\mu^2}\sum_i \frac{(x_i-\mu)^2}{x_i}, \end{aligned}$$

dove $\bar{x}$ indica la media campionaria, $\bar{x} = \frac{1}{n}\sum_i x_i$. Ponendo $l_*(\mu, \lambda) = 0$, risolvendo la prima equazione si ha $\hat{\mu} = \bar{x}$. Quindi, dalla seconda equazione, si ottiene

$$\hat{\lambda} = \frac{n\hat{\mu}^2}{\sum_i \frac{(x_i - \hat{\mu})^2}{x_i}}.$$

Inoltre,

$$\begin{aligned}
\frac{\partial^2 l(\theta)}{\partial \mu^2} &= -\frac{3n\lambda\bar{x}}{\mu^4} + \frac{2n\lambda}{\mu^3} \\
\frac{\partial^2 l(\theta)}{\partial \lambda^2} &= -\frac{n}{2\lambda^2} \\
\frac{\partial^2 l(\theta)}{\partial \mu \partial \lambda} &= \frac{n\bar{x}}{\mu^3} - \frac{n}{\mu^2}.
\end{aligned}$$

Dato che

$$\left.\frac{\partial^2 l(\theta)}{\partial \mu \partial \lambda}\right|_{\hat{\theta}} = 0,$$

si ha che

$$l_{**}(\hat{\theta}) = \begin{pmatrix} -\frac{n\hat{\lambda}}{\hat{\mu}^3} & 0 \\ 0 & -\frac{n}{2\hat{\lambda}^2}. \end{pmatrix}.$$

Ne segue che il punto di stazionarietà trovato è un punto di massimo. Infine, dalla prima identità di Bartlett segue che $E\left[\frac{\partial l(\theta)}{\partial \mu}\right] = 0$, cioè

$$\frac{n\lambda}{\mu^3} E(\bar{X}) = \frac{n\lambda}{\mu^2},$$

da cui $E(\bar{X}) = \mu$. Quindi $\bar{X}$ è stimatore non distorto di μ, che è, date le proprità dello stimatore media campionaria, la media di X.

(c) Siamo sotto condizioni di regolarità, quindi $\hat{\theta} \dot{\sim} N_2(\theta, J^{-1}(\hat{\theta}))$, con

$$J^{-1}(\hat{\theta}) = \begin{pmatrix} \frac{\hat{\mu}^3}{n\hat{\lambda}} & 0 \\ 0 & \frac{2\hat{\lambda}^2}{n}. \end{pmatrix}.$$

(d) L'ipotesi alternativa, da contrapporre all'ipotesi nulla $H_0 : \mu = \mu_0 = 3$, è quella bilaterale, $H_1 : \mu \neq \mu_0$. Sappiamo che $W_P(\mu_0) = 2[l(\hat{\mu}, \hat{\lambda}) - l(\mu_0, \hat{\lambda}_0)]$, con $\hat{\lambda}_0$ stima di massima verosimiglianza vincolata (cioè sotto H_0), e che, sotto H_0, $W_P(\mu_0) \dot{\sim} \chi^2(1)$. Dai dati si ottiene $\hat{\mu} = 2,488$ e $\hat{\lambda} = 1,342$. Inoltre, da

$$l(\mu_0, \lambda) = \frac{n}{2}\log\lambda - \frac{\lambda}{2\mu_0^2}\sum_i x_i + \frac{n\lambda}{\mu_0} - \frac{\lambda}{2}\sum_i \frac{1}{x_i},$$

derivando rispetto a λ, uguagliando a zero e risolvendo si ottiene

$$\hat{\lambda}_0 = \frac{n\mu_0^2}{\sum_i \frac{(x_i - \mu_0)^2}{x_i}} = 1,321.$$

Risulta, quindi, $W_P^{oss} = 0.16$. Dato che la regione critica del test è $\mathcal{R}_\alpha = (\chi^2_\alpha(1), +\infty) = (3,84, +\infty)$ quando $\alpha = 0,05$ e che $W_P^{oss} \notin \mathcal{R}_{0,05}$, H_0 non può essere rifiutata.

(e) Utilizzando la seconda identità di Bartlett, si può scrivere

$$Var\left(\frac{n\lambda}{\mu^3}\bar{X} - \frac{n\lambda}{\mu}\right) = E\left(\frac{3n\lambda}{\mu^4}\bar{X} - \frac{2n\lambda}{\mu^3}\right),$$

da cui

$$\frac{n^2\lambda^2}{\mu^6}Var(\bar{X}) = \frac{3n\lambda\mu}{\mu^4} - \frac{2n\lambda}{\mu^3} = \frac{n\lambda}{\mu^3}.$$

Quindi, posto $\sigma^2 = Var(X)$, si ha $\sigma^2 = \frac{\mu^3}{\lambda}$. Per ottenere un intervallo di confidenza per σ^2, si potrebbe riparametrizzare usando la trasformazione $\left(\mu = \mu, \sigma^2 = \frac{\mu^3}{\lambda}\right)$ con inversa $\left(\mu = \mu, \lambda = \frac{\mu^3}{\sigma^2}\right)$ e usare la quantità pivotale $W_P(\sigma^2)$.

2. Si consideri un campione casuale semplice $y_1, y_2, \ldots, y_n$ da una variabile discreta Y, con supporto $\{0, 1, 3, 5, 7, 9, 11\}$ e funzione di probabilità

$$p(y) = 1 - 1,225\lambda, \quad \text{per} \quad y = 0, \quad \text{e} \quad p(y) = \lambda/(y+1), \quad \text{per} \quad y = 1, 3, 5, 7, 9, 11,$$

dove $\lambda \in (0, \frac{1}{1,225})$ è un parametro ignoto.

(a) Si calcolino media e varianza di Y.

(b) Si ottenga lo stimatore (diciamo $\tilde{\lambda}$) per λ basato sul metodo dei momenti.

(c) Posto che sia $n = 250$ e $\sum_{i=1}^{250} y_i = 75$, si verifichi (ad un livello approssimato del 5%) l'ipotesi che la probabilità dell'evento $\{Y = 0\}$ sia almeno 0,9.

(d) Sapendo che, in dettaglio, l'osservazione campionaria è

f	231	7	4	3	3	1	1
y	0	1	3	5	7	9	11

si ottenga la stima di massima verosimiglianza $\hat{\lambda}$.

Soluzione

(a) Sia $\mathcal{S}_Y$ il supporto di Y. Si ha

$$\begin{aligned} E(Y) &= \sum_{y \in \mathcal{S}_Y} yp(y) = 0 \times p(0) + 1 \times p(1) + 3 \times p(3) + \ldots + 11 \times p(11) \\ &= \lambda\left(\frac{1}{2} + \frac{3}{4} + \frac{5}{6} + \frac{7}{8} + \frac{9}{10} + \frac{11}{12}\right) \\ &= 4,775\lambda, \end{aligned}$$

e

$$\begin{aligned} E(Y^2) &= \sum_{y\in\mathcal{S}_Y} y^2 p(y) = \lambda\left(\frac{1}{2}+\frac{9}{4}+\frac{25}{6}+\frac{49}{8}+\frac{81}{10}+\frac{121}{12}\right) \\ &= 31,225\lambda. \end{aligned}$$

Quindi,

$$Var(Y) = E(Y^2) - [E(Y)]^2 = 31,225\lambda - 22,8\lambda^2.$$

(b) Sia $\bar{y}$ la media campionaria. Dall'equazione $E(Y) = \bar{Y}$, cioè $4,775\lambda = \bar{Y}$, si ricava immediatamente $\tilde{\lambda} = \frac{\bar{Y}}{4,775}$.

(c) Sia $\tau = p(0) = 1 - 1,225\lambda$. L'ipotesi $\tau \geq 0,9$ può essere riscritta come $\lambda \leq \frac{0,1}{1,225} = 0,0816$. Il sistema d'ipotesi può essere dunque scritto come $H_0 : \lambda \leq \lambda_0 = 0,0816$, $H_1 : \lambda > \lambda_0$ e il problema così formulato diventa un problema di verifica d'ipotesi su λ. Utilizzando il Teorema Limite Centrale, possiamo scrivere

$$\tilde{\lambda} \dot{\sim} N\left(\frac{E(Y)}{4,775}, \frac{Var(Y)}{4,775^2 n}\right),$$

cioè

$$\tilde{\lambda} \dot{\sim} N\left(\lambda, \frac{31,225\lambda - 22,8\lambda^2}{250 \times 4,775^2}\right).$$

Per risolvere il problema si può quindi usare la statistica test

$$T = \frac{\tilde{\lambda} - \lambda_0}{\sqrt{\frac{31,225\lambda_0 - 22,8\lambda_0^2}{250\times 4,775^2}}}$$

che, sotto H_0, ha distribuzione approssimata $N(0,1)$. Con i dati, $\tilde{\lambda} = 0,0628$ e $T^{oss} = \frac{-0,0188}{0,02} = -0,94$. Con $\alpha = 0,05$, la regione critica del test è $\mathcal{R}_{0,05} = (1,64, +\infty)$; quindi H_0 non può essere rifiutata.

(d) La funzione di verosimiglianza ha espressione

$$L(\lambda) = \prod_{y\in\mathcal{S}_Y} p(y)^{f(y)} = p(0)^{f(0)} \times p(1)^{f(1)} \times \ldots \times p(11)^{f(11)},$$

in cui $f(y)$ indica la frequenza con la quale il valore y è osservato.

Quindi,

$$\begin{aligned} l(\lambda) &= 231\log(1-1,225\lambda) + 7\log\lambda + 4\log\lambda + 3\log\lambda + 3\log\lambda + \log\lambda + \log\lambda \\ &= 231\log(1-1,225\lambda) + 19\log\lambda. \end{aligned}$$

Derivando rispetto a λ si ottiene

$$l_*(\lambda) = \frac{-1,225\times 231}{1-1,225\lambda} + \frac{19}{\lambda},$$

da cui, uguagliando a zero e risolvendo in λ, si ricava $\hat{\lambda} = 0,062$.

3. In uno studio condotto in Friuli, un campione di 93 maschi adulti, omogenei per stile di vita e non fumatori, è stato classificato secondo il consumo giornaliero di caffè (*non elevato* o *elevato*) e la presenza o meno di patologie coronariche. Tra i 57 soggetti che hanno dichiarato un consumo non elevato di caffè, 15 presentano una patologia coronarica. Lo stesso numero di casi di presenza di una patologia coronarica è stato riscontrato tra i soggetti che hanno dichiarato un consumo elevato di caffè.

(a) Mediante un opportuno test statistico, verificare, ad un livello di significatività del 5%, l'ipotesi che, per la popolazione di riferimento, la probabilità di contrarre una patologia coronarica non dipenda dal livello di assunzione di caffè.

(b) Si calcoli l'intervallo di confidenza al 95% per la probabilità che un soggetto che consumi una elevata quantità di caffè e soffra di patologia coronarica.

Soluzione

(a) I dati possono essere presentati mediate la seguente tabella di contingenza.

	Consumo caffè	
Pat. coronarica	*Non elevato*	*elevato*
Sì	15	15
No	42	21

Si può utilizzare il test X^2 di indipendenza. La tabella delle "frequenze" attese nell'ipotesi H_0 di indipendenza tra presenza di patologia coronarica e livello di assunzione di caffè risulta essere

	Consumo caffè	
Pat. coronarica	*Non elevato*	*elevato*
Sì	18,39	11,61
No	38,61	24,39

Dette O_i e A_i, $i = 1, \ldots, 4$, le frequenze osservate e attese, rispettivamente, la statistica test sarà

$$X^2 = \sum_{i=1}^{4} \frac{(O_i - A_i)^2}{A_i},$$

il cui valore osservato risulta essere $X^2_{oss} = 2,38$. Dato che, sotto H_0, $X^2 \dot{\sim} \chi^2(1)$, la regione critica del test è $\mathcal{R}_\alpha = (\chi^2_\alpha(1), +\infty) = (3,84, +\infty)$ quando $\alpha = 0,05$. I dati, dunque, indicano indipendenza tra presenza di patologia coronarica e livello di assunzione di caffè.

(b) Sia π la probabilità che, nella popolazione di riferimento, un soggetto consumi una elevata quantità di caffè e soffra di una patologia coronarica. La stima naturale di π

è la proporzione campionaria $\hat{\pi} = 15/93 = 0,1613$. L'intervallo di confidenza cercato può essere ottenuto utilizzando la quantità (approssimativamente) pivotale

$$\frac{\hat{\pi} - \pi}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}},$$

e ha per estremi i valori $\hat{\pi} \pm 1,96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$, essendo 1,96 il quantile di ordine 0,975 della distribuzione normale standardizzata. Qui $n = 36$ e l'intervallo risulta essere $(0,0863, 0,2363)$.

4. In un'azienda che produce caffè macinato per moka, viene utilizzata una macchina per il confezionamento di pacchetti dal peso netto nominale di 2,5 etti. Detta X la variabile casuale che descrive la quantità di caffè che la macchina distribuisce in ogni pacchetto, si assume che, quando la macchina funziona correttamente, è $\mu = E(X) = 2,5$, ossia che l'errore commesso è in media nullo.

Si supponga di disporre di un campione (casuale semplice) di 10 pacchetti confezionati, i cui corrispondenti pesi netti $x_1, x_2, \ldots, x_{10}$ risultano essere

$$2,46 \quad 2,53 \quad 2,60 \quad 2,55 \quad 2,49 \quad 2,44 \quad 2,56 \quad 2,47 \quad 2,42 \quad 2,57.$$

(a) Mediante un opportuno tet statistico, si stabilisca se, ad un livello (approssimato) del 5%, tale osservazione campionaria è compatibile con l'ipotesi di corretto funzionamento della macchina.

(b) Assumendo che la macchina stia funzionando correttamente, si usino i dati per costruire un intervallo di confidenza di livello approssimato 0,95 per la varianza, σ^2, di X.

(c) Si supponga ora che sia ragionevole assumere per X una legge normale, $X \sim N(\mu, \sigma^2)$. Si fornisca una regione di confidenza per la coppia (μ, σ^2), di livello approssimato 0,90.

Soluzione

(a) Si tratta di risolvere un problema di verifica d'ipotesi su una singola media. L'ipotesi nulla da considerare è $H_0 : \mu = \mu_0 = 2,5$; ad essa va contrapposta l'ipotesi alternativa $H_1 : \mu \neq \mu_0$. Si può utilizzare la statistica test

$$T = \frac{\hat{\mu} - \mu_0}{\sqrt{\hat{\sigma}^2/n}},$$

con $\hat{\mu} = \bar{x}$, media campionaria, $\hat{\sigma}^2$ varianza campionaria e n dimensione del campione, che sotto H_0 ha distribuzione approssimata dalla legge normale standard. Con $n = 10$,

$\bar{x} = 2,509$ e $\hat{\sigma}^2 = 0,003369$ si ottiene come valore osservato di T, $T_{oss} = 0,4903$. La regione di accettazione del test è $\mathcal{A}_\alpha = (-z_{1-\alpha/2}, z_{1-\alpha/2}) = (1,96, 1,96)$ quando $\alpha = 0,05$, essendo $z_{1-\alpha/2}$ il quantile di ordine $1-\alpha/2$ della distribuzione normale standardizzata. La conclusione, dunque, è che l'osservazione campionaria è compatibile con l'ipotesi di corretto funzionamento della macchina.

(b) Se la macchina funziona correttamente vuol dire che $\mu = 2,5$. Quindi $\hat{\sigma}^2 = E(X^2) - 2,5^2$ e un intervallo di confidenza per $\hat{\sigma}^2$ si può ottenere facilmente da un intervallo per il momento secondo di X. Posto $\gamma = E(X^2)$, lo stimatore naturale per γ è la media dei quadrati

$$\hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Per il Teorema Limite Centrale, $\hat{\gamma} \dot{\sim} N(\gamma,\ \omega/n)$, con $\omega = Var(X^2)$. Inoltre $\hat{\omega} = \frac{1}{n}\sum_{i=1}^{n} X_i^4 - \hat{\gamma}^2$ è stimatore consistente di ω. Quindi, per ottenere un intervallo di confidenza per γ si può usare la quantità (approssimativamente) pivotale

$$\frac{\hat{\gamma} - \gamma}{\sqrt{\hat{\omega}/n}} \dot{\sim} N(0,1) \quad (\text{sotto } \gamma).$$

Dai dati risultano le stime $\hat{\gamma} = 6,298$ e $\hat{\omega} = 39,755 - 6,298^2 = 0,091$ L'intervallo di confidenza per γ, di livello approssimato 0,95, ha per estremi i valori $\hat{\gamma} \pm 1,96\sqrt{\frac{\hat{\omega}}{n}}$ e risulta quindi essere (6,111, 6,485). Di conseguenza, l'intervallo che si ottiene per $\hat{\sigma}^2$ è (0, 0,235).

(c) Se si assume per X un modello normale, la log-verosimiglianza per la coppia (μ, σ^2) ha espressione $l(\mu, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$ e le stime di massima verosimiglianza sono la media campionaria, $\hat{\mu} = \bar{x}$ e la varianza campionaria $\hat{\sigma}^2$. Quindi, dai dati risulta $l(\hat{\mu}, \hat{\sigma}^2) = -\frac{n}{2}(\log\hat{\sigma}^2 + 1) = 23,466$ e, poiché $W(\mu, \sigma^2) \dot{\sim} \chi(2)$ sotto (μ, σ^2), una regione di confidenza per (μ, σ^2), di livello approssimato 0,90, è data da $\{(\mu, \sigma^2) :\ W((\mu, \sigma^2) \leq \chi^2_{0,9}\}$, ossia

$$\left\{(\mu, \sigma^2) :\ l(\mu, \sigma^2) \geq 23,466 - \frac{4,61}{2}\right\}.$$

5. Siano $y_1, y_2, \ldots, y_n$ realizzazioni indipendenti di una variabile casuale Y avente distribuzione continua con densità $p_Y(y;\theta) = 2\left(\frac{\theta}{\pi}\right)^{\frac{1}{2}} e^{-\theta y^2}$, per $y > 0$, con $\theta > 0$ parametro ignoto.

(a) Utilizzando il momento primo di Y, si ottenga lo stimatore $\tilde{\theta}$ basato sul metodo dei momenti. Si mostri che $\tilde{\theta}$ è stimatore consistente di θ.

(b) Si ottenga lo stimatore di massima verosimiglianza $\hat{\theta}$, stabilendo se è non distorto.

(c) Si mostri che, in questo caso, l'equazione di verosimiglianza coincide con l'equazione di stima derivante dall'applicazione del metodo dei momenti e l'uso del momento secondo di Y.

(d) Si forniscano opportune approssimazioni per le funzioni di ripartizione di $\hat{\theta}$ e $\tilde{\theta}$, cioè, approssimazioni di $\Pr\{\hat{\theta} \leq c\}$ e $\Pr\{\tilde{\theta} \leq c\}$, dove $c > 0$ è un valore fissato.

(e) Si supponga che sia $n = 10$ e che l'osservazione campionaria sia tale che $\sum_i y_i^2 = 3,5$. Si risolva, ad un livello di significatività (approssimato) 0,05, il problema di verifica d'ipotesi $H_0 : \theta = 1$ contro $H_1 : \theta > 1$.

Soluzione

(a) Dato che $\frac{d\ e^{-\theta y^2}}{dy} = -2\theta y e^{-\theta y^2}$, si ha

$$E(Y) = 2\left(\frac{\theta}{\pi}\right)^{\frac{1}{2}} \int_0^{+\infty} y e^{-\theta y^2} dy = 2\left(\frac{\theta}{\pi}\right)^{\frac{1}{2}} \left[-\frac{e^{-\theta y^2}}{2\theta}\right]_0^{+\infty} = 2\left(\frac{\theta}{\pi}\right)^{\frac{1}{2}} \left[0 + \frac{1}{2\theta}\right] = \frac{1}{\sqrt{\pi\theta}}.$$

Pertanto, dall'equazione $\bar{Y} = \frac{1}{\sqrt{\pi\theta}}$, dove $\bar{Y}$ è la media campionaria, si ricava $\tilde{\theta} = \frac{1}{\pi \bar{Y}^2}$. Per la legge dei grandi numeri, $\bar{Y} \xrightarrow{p} E(Y)$. Dato che $\tilde{\theta}$ è funzione continua di $\bar{Y}$ si ha che $\tilde{\theta} \xrightarrow{p} \frac{1}{\pi E(Y)^2} = \theta$. Quindi, $\tilde{\theta}$ è stimatore consistente di θ.

(b) La funzione di verosimiglianza per θ è

$$L(\theta) = 2^n \left(\frac{\theta}{\pi}\right)^{\frac{n}{2}} \prod_{i=1}^{n} e^{-\theta y_i^2} \ \propto\ \theta^{\frac{n}{2}} e^{-\theta \sum_{i=1}^n y_i^2}.$$

Passando al logaritmo si ha, $l(\theta) = \frac{n}{2}\log\theta - \theta\sum_{i=1}^n y_i^2$, e derivando rispetto a θ si ottiene la funzione punteggio $l_*(\theta) = \frac{n}{2\theta} - \sum_{i=1}^n y_i^2$. Risolvendo l'equazione di verosimiglianza $l_*(\theta) = 0$, si ricava $\hat{\theta} = \frac{n}{2\sum_i y_i^2}$. Inoltre, $l_{**}(\theta) = \frac{d\ l_*(\theta)}{d\theta} = -\frac{n}{2\theta^2} < 0\ \forall\theta$; quindi, la radice dell'equazione di verosimiglianza è punto di massimo globale. Ora, dalla prima identità di Bartlett, $E[l_*(\theta)] = 0$, si ricava che $E(\sum_i Y_i^2) = \frac{n}{2\theta}$, da cui $E(Y^2) = \frac{1}{2\theta}$. Inoltre, $E\left(\frac{\sum_i Y_i^2}{n}\right) = E(Y^2)$ e, poiché $\hat{\theta}$ è funzione convessa di $\frac{\sum_i Y_i^2}{n}$, dalla disuguaglianza di Jensen risulta che $E(\hat{\theta}) > \frac{1}{2E(Y^2)} = \theta$. Quindi $\hat{\theta}$ è stimatore distorto di θ.

(c) L'equazione di stima derivante dall'uso del metodo dei momenti con momento secondo di Y è $E(Y^2) = \frac{\sum_i y_i^2}{n}$, ossia, $\frac{1}{2\theta} = \frac{\sum_i y_i^2}{n}$, che coincide, evidentemente, con l'equazione di verosimiglianza.

(d) Sappiamo che, sotto condizioni di regolarità, $\hat{\theta} \dot{\sim} N(\theta, i(\theta)^{-1})$. Nel nostro caso, $i(\theta) = \frac{n}{2\theta^2}$ e quindi, per $c > 0$,

$$\Pr\{\hat{\theta} \leq c\} = \Pr\left\{\frac{\hat{\theta} - \theta}{\sqrt{\frac{2\theta^2}{n}}} \leq \frac{c - \theta}{\sqrt{\frac{2\theta^2}{n}}}\right\} \doteq \Phi\left(\frac{c - \theta}{\sqrt{\frac{2\theta^2}{n}}}\right),$$

dove $\Phi(\cdot)$ indica la funzione di ripartizione della normale standard. Per quanta riguarda lo stimatore $\tilde{\theta}$, abbiamo

$$\begin{aligned}\Pr\{\tilde{\theta} \leq c\} &= \Pr\left\{\frac{1}{\pi \bar{Y}^2} \leq c\right\} = \Pr\left\{\bar{Y}^2 \geq \frac{1}{\pi c}\right\} = \Pr\left\{\bar{Y} \geq \sqrt{\frac{1}{\pi c}}\right\} \\ &\doteq \left[1 - \Phi\left(\frac{\sqrt{\frac{1}{\pi c}} - \sqrt{\frac{1}{\pi\theta}}}{\sqrt{\frac{1}{n}\left(\frac{1}{2\theta} - \frac{1}{\pi\theta}\right)}}\right)\right],\end{aligned}$$

dato che $\bar{Y} \dot{\sim} N(\mu, \sigma^2/n)$ per il Teorema Limite Centrale, con $\mu = E(Y) = \sqrt{\frac{1}{\pi\theta}}$ e $\sigma^2 = Var(Y) = E(Y^2) - E(Y)^2 = \frac{1}{2\theta} - \frac{1}{\pi\theta}$.

(e) Si può utilizzare la statistica test

$$r_e(\theta_0) = \frac{\hat{\theta} - \theta_0}{\sqrt{i(\theta_0)^{-1}}} = \frac{\hat{\theta} - \theta_0}{\sqrt{\frac{2\theta_0^2}{n}}},$$

il cui valore osservato è

$$r_e^{oss} = \frac{1,4286 - 1}{\sqrt{0,2}} = 0,958.$$

Dato che la regione critica di livello approssimato $\alpha = 0,05$ è $\mathcal{R}_\alpha = (z_{1-\alpha}, +\infty) = (1,64, +\infty)$, l'ipotesi nulla non può essere rifiutata.

6. Nelle elezioni politiche del 4 marzo 2018 ha votato il 72,9% degli Italiani aventi diritto. I risultati (in termini percentuali) sono riassunti dalla tabella che segue.

M5S	Area CD	Area CS	Altri
32,7	37	24,6	5,7

Una settimana dopo il voto, una nota testata giornalistica ha commissionato ad un ente di ricerca un sondaggio sull'opinione di quanti non si erano recati alle urne. Così, in una intervista, ad un campione di 1250 soggetti, rappresentativo degli Italiani che non avevano votato, è stato chiesto di esprimere una preferenza. Il risultato del sondaggio è stato di 225 preferenze espresse per il M5S, 482 per l'Area CD, 455 per l'Area CS e 88 per Altri.

(a) Mediante un test opportuno, si stabilisca se il risultato del sondaggio è coerente con il risultato elettorale del 4 marzo.

(b) Sulla base del risultato del sondaggio, è ragionevole affermare che tra gli Italiani che non hanno votato il 4 marzo la quota di quelli schierati per l'Area CD sia la stessa di quelli schierati per l'Area CS?

Soluzione

(a) È ragionevole ritenere che l'osservazione campionaria (225, 482, 455, 88) sia realizzazione di una variabile casuale multinomiale a 4 celle, di indice 1250 e vettore di parametri $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)$. Per rispondere al quesito bisogna risolvere il problema di verifica d'ipotesi in cui $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi_0}$ e $H_1 : \overline{H}_0$, con $\boldsymbol{\pi_0} = (\pi_{10}, \pi_{20}, \pi_{30}, \pi_{40}) = (0,327,\ 0,37,\ 0,246,\ 0,057)$. Si può utilizzare la statistica test rapporto di verosimiglianza $W(\boldsymbol{\pi_0}) = 2[l(\hat{\boldsymbol{\pi}}) - l(\boldsymbol{\pi_0})]$, che sotto H_0 ha distribuzione approssimata $\chi^2(3)$. Abbiamo che

$$l(\boldsymbol{\pi}) = 225 \log \pi_1 + 482 \log \pi_2 + 455 \log \pi_3 + 88 \log \pi_4,$$

con $\pi_4 = 1 - \pi_1 - \pi_2 - \pi_3$, e sappiamo che le stime di massima verosimiglianza sono $\hat{\pi}_1 = 225/1250 = 0,18$, $\hat{\pi}_2 = 482/1250 = 0,3856$, $\hat{\pi}_3 = 455/1250 = 0,364$, $\hat{\pi}_4 = 88/1250 = 0,0704$. Quindi risulta che $W(\boldsymbol{\pi_0}) = 2[-1538,491 + 1620,93] = 164,878$. Fissando un livello di significatività nominale $\alpha = 0,05$, la regione critica del test è $\mathcal{R}_\alpha = (\chi^2_{1-\alpha}(3), +\infty) = (7,81, +\infty)$ e l'ipotesi nulla viene evidentemente rifiutata: il risultato del sondaggio non è coerente con il risultato elettorale del 4 marzo.

(b) Anche in questo caso il problema da risolvere è un problema di verifica d'ipotesi in cui l'ipotesi nulla può essere formulata come $H_0 : \pi_2 = \pi_3$. L'alternativa da contrapporre è ancora $H_1 : \overline{H}_0$. La statistica test da utilizzare questa volta è

$$W_P^{H_0} = 2[l(\hat{\boldsymbol{\pi}}) - l(\hat{\boldsymbol{\pi}}_{\boldsymbol{0}})],$$

dove $\hat{\boldsymbol{\pi}}_{\boldsymbol{0}}$ indica la stima di massima verosimiglianza vincolata alla validità dell'ipotesi nulla. Sotto H_0, $W_P^{H_0} \dot{\sim} \chi^2(k)$, dove k è la differenza tra la dimensione dello spazio parametrico non vincolato e quello vincolato. In questo caso $k = 3 - 2 = 1$. La funzione di log-verosimiglianza vincolata all'ipotesi $\pi_2 = \pi_3 = \pi^*$ ha espressione

$$l_P^{H_0} = 225 \log \pi_1 + (482 + 455) \log \pi^* + 88 \log \pi_4,$$

con $\pi_4 = 1 - \pi_1 - 2\pi^*$. Derivando rispetto a π_1 e π^* e risolvendo il sistema di equazioni di verosimglianza, si ottengono le stime $\hat{\pi}_1 = 225/1250 = 0,18$, $\hat{\pi}^* = (482+455)/2500 = 0,3748$ e $\hat{\pi}_4 = 88/1250 = 0,0704$. Quindi Il valore osservato della statistica test è $W_P^{H_0} = 2[-1538,491 + 1538.88] = 0,778$. Dato che la regione critica di livello approssimato $\alpha = 0,05$ è $\mathcal{R}_\alpha = (\chi^2_{1-\alpha}(1), +\infty) = (3,84, +\infty)$, l'ipotesi nulla non può essere rifiutata: sulla base del risultato del sondaggio, è ragionevole affermare che tra gli Italiani che non hanno votato il 4 marzo la quota di quelli schierati per l'Area CD è la stessa di quelli schierati per l'Area CS.

7. Sia X la variabile casuale che descrive il numero di clienti che usano un certo sportello bancomat dalle 8 alle 12 di ogni mercoledì. Si assume che X abbia funzione di probabilità $p_X(x) = \theta_1^x (1+\theta_1)^{-(x+1)}$, per $x \in \{0, 1, 2, 3, ...\}$, con $\theta_1 > 0$ parametro ignoto. Lo sportello viene monitorato per n settimane e i dati raccolti, $x_1, x_2, \ldots, x_n$, si suppongono costituire un campione casuale semplice da X.

(a) Si ottenga lo stimatore di massima verosimiglianza $\hat{\theta}_1$.

(b) Si stabilisca se $\hat{\theta}_1$ coincide con lo stimatore basato sul metodo dei momenti (con l'uso del momento primo di X).

(c) Si ottenga un'espressione per la varianza di X e se ne fornisca uno stimatore.

Sia ora Y la variabile che descrive il numero di clienti che usano lo sportello dalle 14 alle 18 (sempre di mercoledì). Sia $y_1, y_2, \ldots, y_n$, l'osservazione (campione casuale semplice) relativa a Y. Si assume che Y sia indipendente da X e che abbia legge di Poisson, cioè, che abbia funzione di probabilità $p_Y(y) = e^{-\theta_2}\theta_2^y(1/y!)$, per $y \in \{0, 1, 2, 3, ...\}$, con $\theta_2 > 0$ parametro ignoto.

(d) Si scriva la funzione di log-verosimiglianza per la coppia (θ_1, θ_2).

(e) Supponendo che sia $n = 12$ e che l'osservazione campionaria sia tale che $\sum_{i=1}^n x_i = 533$ e $\sum_{i=1}^n y_i = 467$, si risolva, mediante l'uso di una statistica test basata sul rapporto di verosimiglianza, il problema di verifica d'ipotesi $H_0 : \theta_1 = \theta_2$ contro $H_1 : \theta_1 \neq \theta_2$.

Soluzione

(a) Si ha $L(\theta_1) = \prod_{i=1}^n \theta_1^{x_i}(1+\theta_1)^{-(x_i+1)}$, da cui, passando al logaritmo, $l(\theta_1) = \log\theta_1 \sum_{i=1}^n x_i - \log(1+\theta_1)\sum_{i=1}^n (x_i+1)$. Derivando una e due volte rispetto a θ_1, si ottengono

$$l_*(\theta_1) = \frac{1}{\theta_1}\sum_{i=1}^n x_i - \frac{1}{1+\theta_1}\left(\sum_{i=1}^n x_i + n\right)$$

e

$$l_{**}(\theta_1) = -\frac{1}{\theta_1^2}\sum_{i=1}^n x_i + \frac{1}{(1+\theta_1)^2}\left(\sum_{i=1}^n x_i + n\right).$$

Risolvendo l'equazione di verosimiglianza, $l_*(\theta_1) = 0$, si ottiene immediatamente

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}.$$

Inoltre, $l_{**}(\hat{\theta}_1) = -\frac{n}{\bar{x}} + \frac{n}{1+\bar{x}} < 0$, purché sia $\bar{x} > 0$.

(b) Dalla prima identità di Bartlett, $E[l_*(\theta_1)] = 0$, si ottiene

$$\frac{nE(X)}{\theta_1} - \frac{nE(X)+n}{1+\theta_1},$$

da cui $E(X) = \theta_1$. Ne segue che lo stimatore risultante dall'applicazione del metodo dei momenti (con uso del momento primo) coincide con la media campionaria $\bar{X}$.

(c) Dalla seconda identità di Bartlett, $Var[l_*(\theta_1)] = E[-l_{**}(\theta_1)]$, si ottiene

$$\frac{\sum_i Var(X_i)}{\theta_1^2(1+\theta_1)^2} = \frac{\sum_i E(X_i)}{\theta_1^2} - \frac{\sum_i E(X_i)+n}{(1+\theta_1)^2} = \frac{n}{\theta_1(1+\theta_1)},$$

da cui

$$Var(X) = \theta_1(1+\theta_1).$$

Uno stimatore per $Var(X)$ si può ottenere sostituendo $\hat{\theta}_1$ a θ_1 nella espressione sopra riportata.

(d) In questo caso,

$$L(\theta_1, \theta_2) \propto \prod_{i=1}^{n} \theta_1^{x_i}(1+\theta_1)^{-(x_i+1)} e^{-\theta_2} \theta_2^{y_i}$$

e quindi

$$l(\theta_1, \theta_2) = \log\theta_1 \sum_{i=1}^{n} x_i - \log(1+\theta_1) \sum_{i=1}^{n}(x_i+1) - n\theta_2 + \log\theta_2 \sum_{i=1}^{n} y_i.$$

(e) La statistica test da utilizzare è $W_P^{H_0} = 2[l(\hat{\theta}_1, \hat{\theta}_2) - l(\hat{\theta}_0, \hat{\theta}_0)]$, dove $\hat{\theta}_0$ è la stima di massima verosimiglianza del valore comune di θ_1 e θ_2 sotto H_0. Per determinare $\hat{\theta}_0$, bisogna considerare la log-verosimiglianza parziale

$$l_P(\theta_0) = l(\theta_0, \theta_0) = \log\theta_0 \sum_{i=1}^{n} x_i - \log(1+\theta_0) \sum_{i=1}^{n}(x_i+1) - n\theta_0 + \log\theta_0 \sum_{i=1}^{n} y_i,$$

da cui si ottiene la funzione punteggio

$$l_{P*}(\theta_0) = \frac{\sum_i x_i}{\theta_0} - \frac{\sum_i x_i + n}{1+\theta_0} - n + \frac{\sum_i y_i}{\theta_0}.$$

Ponendo pari a zero si ricava l'equazione di stima

$$\sum_i x_i - 2n\theta_0 - n\theta_0^2 + \theta_0 \sum_i y_i + \sum_i y_i = 0,$$

che si può riscrivere come

$$\theta_0^2 + (2-\bar{y})\theta_0 - (\bar{x}+\bar{y}) = 0.$$

Sotto la stessa condizione fissata al punto (a), cioè $\bar{x} > 0$, questa equazione ha un'unica radice ammissibile e risulta

$$\hat{\theta}_0 = \frac{-(2-\bar{y}) + \sqrt{(2-\bar{y})^2 + 4(\bar{x}+\bar{y})}}{2}.$$

Con i dati, $\hat{\theta}_0 = 39,050$, $\hat{\theta}_1 = 44,416$, $\hat{\theta}_2 = \bar{y} = 38,916$, $l(\hat{\theta}_1, \hat{\theta}_2) = 1185,227$, $l_{P*}(\hat{\theta}_0) = 1185,123$ e $W_P^{H_0} = 0,208$. Dato che sotto H_0 $W_P^{H_0} \dot{\sim} \chi^2(1)$, ad un livello di significatività (approssimato) $\alpha = 0.05$, l'ipotesi nulla non può essere rifiutata.

8. Siano $y_1, y_2, \ldots, y_n$ realizzazioni indipendenti di una variabile casuale Y avente distribuzione continua con densità $p_Y(y;\theta,\lambda) = \lambda e^{-\lambda(y-\theta)}$, per $y \geq \theta$, con $\theta > 0$ e $\lambda > 0$ parametri ignoti.

(a) Si identifichino lo spazio campionario e lo spazio parametrico e si stabililisca se il modello considerato ha verosimiglianza regolare.

(b) Si scriva la funzione di verosimiglianza per (θ, λ), individuando una statistica sufficiente.

(c) Si ottenga lo stimatore di massima verosimiglianza $(\hat{\theta}, \hat{\lambda})$.

(d) Si ottenga la funzione di ripartizione di Y, cioè la $\Pr\{Y \leq c\}$, con $c > \theta$ reale fissato. Si stabilisca se $\hat{\theta}$ è stimatore consistente di θ.

Soluzione

(a) Lo spazio campionario è $\mathcal{Y} = [\theta, +\infty)^n$, mentre lo spazio parametrico è $(0, +\infty) \times (0, +\infty)$. Il modello non ha verosimiglinaza regolare, visto che lo spazio campionario dipende da parte del parametro e la funzione di verosimiglianza non può essere derivabile ovunque.

(b) Si ha

$$\begin{aligned} L(\theta,\lambda) &= \prod_{i=1}^{n} \lambda e^{-\lambda(y_i-\theta)} I_{[\theta,+\infty)}(y_i) = \lambda^n e^{-\lambda \sum_i (y_i-\theta)} \prod_{i=1}^{n} I_{[\theta,+\infty)}(y_i) \\ &= \lambda^n e^{-n\lambda\bar{y}} e^{n\lambda\theta} I_{[\theta,+\infty)}(y_{(1)}), \end{aligned}$$

in cui $\bar{y}$ è la media campionaria e $y_{(1)}$ è la più piccola osservazione nel campione. Evidentemente, una statistica sufficiente per l'inferenza su (θ, λ) è la coppia $(Y_{(1)}, \sum_i Y_i)$.

(c) Per λ fissato, $L(\theta;\lambda) = h(\lambda) e^{n\lambda\theta} I_{(0,y_{(1)}]}(\theta)$ che è funzione strettamente crescente di θ in $(0, y_{(1)}]$. Quindi, $\hat{\theta}(\lambda) = \hat{\theta} = Y_{(1)}$. Di conseguenza, $L(\lambda;\hat{\theta}) = \lambda^n e^{-n\lambda\bar{y}} e^{n\lambda y_{(1)}}$, da cui $l(\lambda;\hat{\theta}) = n \log\lambda - n\lambda\bar{y} + n\lambda y_{(1)}$ e

$$l_*(\lambda;\hat{\theta}) = \frac{n}{\lambda} - n\bar{y} + n y_{(1)}.$$

Risolvendo l'equazione di verosimiglianza in λ si ottiene

$$\hat{\lambda} = \frac{1}{\bar{Y} - Y_{(1)}}.$$

(d) Per $c > \theta$ fissato, si ha

$$\Pr\{Y \leq c\} = \int_{\theta}^{c} \lambda e^{-\lambda(y-\theta)} dy = \int_{0}^{c-\theta} \lambda e^{-\lambda t} dt$$

che è la funzione di ripartizione di una variabile casuale esponenziale di parametro λ calcolata in $c-\theta$. Quindi, $\Pr\{Y \le c\} = 1 - e^{-\lambda(c-\theta)}$. Ora, $\hat{\theta} = Y_{(1)}$ e, per ogni $\epsilon > 0$ fissato,

$$\Pr\{Y_{(1)} \le \theta + \epsilon\} = 1 - \Pr\{Y_{(1)} > \theta + \epsilon\} = 1 - \prod_{i=1}^{n} \Pr\{Y_i > \theta + \epsilon\} = 1 - \left(e^{-\lambda\epsilon}\right)^n$$

tende a zero quando n cresce (per ogni $\lambda > 0$). Pertanto, $\hat{\theta} \xrightarrow{p} \theta$.

9. È ragionevole assumere che il tempo Y (in mesi) di tracciabilità nel plasma (dal momento della somministrazione) di una sostanza farmacologica segua una legge con funzione di densità $p(y;\tau) = 2\tau y e^{-y^2}(1-e^{-y^2})^{\tau-1}$, per $y > 0$ e con $\tau > 0$ parametro ignoto. Sia $y_1, y_2, \ldots, y_n$, un campione del tempo di tracciabilità raccolto su n cavie.

(a) Si determini lo stimatore di massima verosimiglianza $\hat{\tau}$, stabilendo se è non distorto.

(b) Si mostri che $V = -\log(1 - e^{-Y^2})$ ha distribuzione esponenziale di parametro τ.

(c) Con n=5 e con un'osservazione campionaria tale che $\sum_{i=1}^{20} \log(1 - e^{-y_i^2}) = -4,25$, si stabilisca, ad un livello di significatività esatto del 5%, se è ragionevole ipotizzare che sia $\tau > 1$.

(d) Si risolva (ad un livello approssimato) il problema di cui al punto (c), utilizzando i risultati della teoria asintotica relativi a $\hat{\tau}$.

Soluzione

(a) La funzione di verosimiglianza risulta essere:

$$L(\tau) = \prod_{i=1}^{n} p(y_i;\tau) = \prod_{i=1}^{n} 2\tau y_i e^{-y_i^2}(1 - e^{-y_i^2})^{\tau-1} \propto \tau^n \prod_{i=1}^{n} (1 - e^{-y_i^2})^{\tau}.$$

Di conseguenza, log verosimiglianza e funzione punteggio hanno la seguente forma:

$$l(\tau) = n \log \tau + \tau \sum_{i=1}^{n} \log(1 - e^{-y_i^2}) \quad \text{e} \quad l_*(\tau) = n/\tau + \sum_{i=1}^{n} \log(1 - e^{-y_i^2}).$$

Dall'equazione di verosimiglianza, $l_*(\tau) = 0$, si ottiene

$$\hat{\tau} = \frac{n}{-\sum_{i=1}^{n} \log(1 - e^{-y_i^2})},$$

e, poiché $l_{**}(\tau) = -n/\tau^2 < 0$, $\hat{\tau}$ è stima di massima verosimiglianza. Dalla prima identità di Bartlett si ricava facilmente che

$$E\left[-\sum_{i=1}^{n} \log(1 - e^{-Y_i^2})\right] = \frac{n}{\tau}.$$

Allora, sfruttando la diseguaglianza di Jensen,

$$E(\hat{\tau}) = E\left[\frac{n}{-\sum_{i=1}^{n}\log(1-e^{-Y_i^2})}\right] < \frac{n}{E\left[-\sum_{i=1}^{n}\log(1-e^{-Y_i^2})\right]} = \tau.$$

Quindi $\hat{\tau}$ è stimatore distorto di τ.

(b) Si ha che $y \in (0,+\infty)$, $y^2 \in (0,+\infty)$, $e^{-y^2} \in (0,1)$, $1-e^{-y^2} \in (0,1)$ e $-\log(1-e^{-y^2}) \in (0,+\infty)$. Inoltre, se $v = -\log(1-e^{-y^2})$, allora risulta essere $y = [-\log(1-e^{-v})]^{1/2}$. Infine,

$$\frac{dy(v)}{dv} = \frac{1}{2}\left[-\log(1-e^{-v})\right]^{-1/2}\left[-\frac{e^{-v}}{1-e^{-v}}\right].$$

Di conseguenza,

$$\begin{aligned} p_V(v;\tau) = p_Y(y(v);\tau)\left|\frac{dy(v)}{dv}\right| &= 2\tau\left[-\log(1-e^{-v})\right]^{1/2}(1-e^{-v})(e^{-v})^{\tau-1} \\ &\times \frac{1}{2}\left[-\log(1-e^{-v})\right]^{-1/2}\frac{e^{-v}}{1-e^{-v}} \\ &= \tau e^{-\tau v}, \end{aligned} \tag{1}$$

per $v > 0$. Quindi V ha distribuzione esponenziale di parametro τ.

(c) Abbiamo visto che lo stimatore di massima verosimiglianza è $\hat{\tau} = n/\left[-\sum_{i=1}^{n}\log(1-e^{-Y_i^2})\right]$. Dato che $\sum_{i=1}^{n} -\log(1-e^{-Y_i^2}) \sim Ga(n,\tau)$, si ha che $n/\hat{\tau} \sim Ga(n,\tau)$ e, quindi,

$$\frac{2n\tau}{\hat{\tau}} \sim Ga(n,1/2) \quad \text{ovvero} \quad \sim \chi^2(2n).$$

Allora, posto $H_0 : \tau = 1 = \tau_0$ e $H_1 : \tau > 1$, si può usare la statistica test $T = 2n\tau_0/\hat{\tau} = 2n/\hat{\tau}$, che sotto H_0 ha distribuzione $\chi^2(2n)$. Per come è fatta la statistica (e per la forma di H_1), sono i valori piccoli ad essere contrari ad H_0. Quindi, con $n = 5$ la regione critica, di livello esatto 0,05, è $R_{0,05} = (0, \chi^2_{0,05}(10)) = (0,\ 3,94)$. Dai dati, $\hat{\tau} = 1,176$ e $T^{oss} = 8,5$. Di conseguenza, l'ipotesi nulla non può essere rifiutata .

(d) Dalla teoria asintotica è noto che

$$\frac{\hat{\tau}-\tau}{\sqrt{j^{-1}(\hat{\tau})}} \dot{\sim} N(0,1)$$

sotto τ. In questo caso $j^{-1}(\tau) = \tau^2/n$ e, con $n = 5$, la statistica test da utilizzare è

$$r_e(\tau_0) = \frac{\hat{\tau}-\tau_0}{\hat{\tau}/\sqrt{5}},$$

con $\tau_0 = 1$. Il valore osservato è $r_e^{oss} = 0,3346$ e, essendo la regione critica (di livello approssimato) $R_{0,05} = (z_{0,95}, +\infty) = (1,64, +\infty)$, l'ipotesi nulla non viene rifiutata.

10. Di recente, ad un gruppo di 81 studenti frequentanti il corso di Statistica 2 è stato somministrato un questionario. Ogni studente ha fornito, tra l'altro, la propria valutazione (su scala $1-10$) della *continuità nell'impegno allo studio durante il corso*. I risultati sono raccolti nella tabella seguente.

Valutazione	≤ 4	$5-7$	$8-10$
frequenza	20	43	18

Si assuma che il gruppo di studenti in questione sia un campione rappresentativo della popolazione di tutti gli studenti che hanno frequentato il corso negli ultimi 3 anni.

(a) Mediante un opportuno test statistico, verificare l'ipotesi che, nella popolazione di riferimento, la probabilità che uno studente valuti il suo impegno con punteggio ≤ 7 sia almeno pari a 0,8.

(b) Si calcoli la potenza del test di cui al punto precedente quando il parametro di interesse vale 0,7.

11. Sia X la variabile casuale che descrive un certo carattere in una certa popolazione. Si suppone che la legge di X sia un elemento della famiglia di distribuzioni uniformi, $X \sim U[\mu - \sqrt{3}\gamma, \mu + \sqrt{3}\gamma]$, con μ reale e $\gamma > 0$ parametri ignoti. Si supponga di disporre di un campione casuale semplice $x_1, x_2, \ldots, x_n$ da X.

(a) Si ottenga lo stimatore $(\tilde{\mu}, \tilde{\gamma})$ basato sul metodo dei momenti, stabilendo se è non distorto.

(b) Supponendo che l'osservazione campionaria sia tale che $n = 5$, $\sum_{i=1}^{n} x_i = 7,85$ e $\sum_{i=1}^{n} x_i^2 = 24,55$, si risolva ad un livello approssimato del 5% il problema di verifica d'ipotesi $H_0 : \mu \geq 2$ contro $H_1 : \mu < 2$.

(c) Si calcoli la potenza del test di cui al punto precedente, in corrispondenza dell'alternativa $\mu = 1$.

(d) Posto $\mu = 0$, si ottenga lo stimatore di massima verosimiglianza $\hat{\gamma}$.

(e) Si fornisca la distribuzione di $\hat{\gamma}$.

12. Siano $y_1, y_2, \ldots, y_n$ realizzazioni indipendenti di una variabile casuale Y avente distribuzione continua con densità $p_Y(y;\theta) = \theta(1+y)^{-(\theta+1)}$, per $y > 0$, con $\theta > 0$ parametro ignoto.

(a) Si ottenga lo stimatore di massima verosimiglianza $\hat{\theta}$.

(b) Si mostri che $\hat{\theta}$ è stimatore consistente di θ.

(c) Si mostri che il valore atteso $\mu = E(Y)$ è finito solo se $\theta > 1$ e vale $\mu = \frac{1}{\theta-1}$.

(d) Si assuma che sia $\theta > 1$. Usando la quantità pivotale $\frac{2n\theta}{\hat{\theta}}$, si ottenga un intervallo di confidenza per μ di livello esatto 0,95.

13. La tabella che segue riporta i valori della funzione di log-verosimiglianza (cambiata di segno) calcolata per un campione casuale semplice di dimensione $n = 20$ e relativa ad un modello statistico indicizzato da una coppia di parametri (β, γ), i cui possibili valori sono quelli riportati, rispettivamente, nell'ultima riga e nella prima colonna (da sinistra) della tabella stessa.

$$-\ell(\beta, \gamma)$$

	5	42,130	39,479	39,593	45,060	52,557	58,028	84,451
	3	35,558	34,686	36,215	41,579	46,983	49,962	56,948
	1	35,543	34,689	36,220	41,538	46,851	49,751	56,523
γ	0	37,615	37,064	37,816	39,770	41,014	41,579	44,148
	-1	42,140	41,180	40,035	38,395	37,649	37,563	39,647
	-3	57,668	52,534	43,935	37,064	34,984	34,768	37,806
	-5	63,998	55,735	44,784	37,233	35,160	34,956	38,754
		-5	-3	-1	0	1	3	5
					β			

Si supponga siano validi i risultati della teoria asintotica.

(a) Si fornisca la stima di massima verosimiglianza di (β, γ).

(b) Si stabilisca, motivando, se l'insieme $\{(3, -5)\ (1, -5)\ (1, -3)\ (3, -3)\}$ costituisce una regione di confidenza (di un qualche livello approssimato) per (β, γ), basata su $W(\beta, \gamma)$. Si individui la regione di confidenza di livello approssimato 0,95.

(c) Si risolva, ad un livello di significatività approssimato $\alpha = 0,05$, il problema di verifica d'ipotesi $H_0 : \beta = 0$ contro $H_1 : \beta < 0$.