

Università degli studi di Padova  
Dipartimento di Scienze Statistiche  
Corso di Laurea Triennale in  
Statistica, Economia e Finanza



RELAZIONE FINALE

**INFERENZA BASATA SULLA  
VEROSIMIGLIANZA PER L'INDICE DI YAUDEN**

Relatore Prof. Nicola Sartori  
Dipartimento di Scienze Statistiche

Laureanda Ainhoa-Elena Léger  
Matricola N 1051560

Anno Accademico 2015/2016



# Indice

<b>Introduzione</b>	<b>5</b>
<b>1 Test diagnostici e indice di Youden</b>	<b>7</b>
1.1 Test diagnostici . . . . .	7
1.2 Sensibilità, specificità e curva ROC . . . . .	9
1.3 Indice di Youden . . . . .	11
1.4 Stima dell'indice di Youden . . . . .	13
<b>2 L'inferenza di verosimiglianza</b>	<b>15</b>
2.1 Introduzione . . . . .	15
2.2 Specificazione del modello . . . . .	16
2.3 Verosimiglianza . . . . .	16
2.3.1 Funzione di verosimiglianza . . . . .	16
2.3.2 Stima di massima verosimiglianza . . . . .	18
2.3.3 Statistiche sufficienti . . . . .	19
2.4 Parametri di disturbo . . . . .	21
2.5 Verosimiglianza profilo . . . . .	22
2.6 Invarianza rispetto a riparametrizzazioni . . . . .	24
2.7 Inferenza di verosimiglianza . . . . .	25
2.7.1 Test d'ipotesi . . . . .	26
2.7.2 Stima intervallare . . . . .	29
<b>3 Teoria asintotica di ordine superiore</b>	<b>31</b>
3.1 Introduzione . . . . .	31
3.2 La formula $p^*$ . . . . .	32
3.3 La statistica radice con segno modificata $r^*$ . . . . .	33

3.4	Approssimazione di Skovgaard . . . . .	35
3.5	Pacchetto <code>LikelihoodAsy</code> . . . . .	36
<b>4</b>	<b>Studi di simulazione</b>	<b>39</b>
4.1	Introduzione . . . . .	39
4.2	Procedura di confronto fra $r$ e $r^*$ . . . . .	40
4.3	Distribuzione normale . . . . .	41
4.3.1	Normali con varianze uguali . . . . .	42
4.3.2	Normali con varianze diverse . . . . .	47
4.4	Distribuzione gamma . . . . .	54
	<b>Conclusione</b>	<b>61</b>
	<b>Bibliografia</b>	<b>62</b>
<b>A</b>	<b>Codice R</b>	<b>65</b>
A.1	Caso normale con varianze uguali . . . . .	65
A.2	Caso normale con varianze diverse . . . . .	68
A.3	Caso gamma . . . . .	72

# Introduzione

La funzione di verosimiglianza rappresenta la base di molti metodi statistici inferenziali per risolvere problemi di stima, in particolare la costruzione di intervalli di confidenza, e per realizzare test statistici. Queste procedure si basano su approssimazioni asintotiche dette del primo ordine, derivanti dall'applicazione del teorema limite centrale, come ad esempio la normalità della distribuzione campionaria dello stimatore di massima verosimiglianza o la distribuzione chi-quadrato del log-rapporto di verosimiglianza.

Tuttavia queste approssimazioni possono condurre a risultati insoddisfacenti nel caso di campioni di dimensione ridotta o nel caso di scarsa informazione sui parametri, dovuta alla presenza di più parametri di disturbo. Per ovviare a questi problemi, negli anni '80 dello scorso secolo si è sviluppata la teoria asintotica di ordine superiore, con metodi che migliorano la precisione delle approssimazioni distributive delle quantità di verosimiglianza.

In questa tesi ci si interessa in particolare alla statistica  $r^*$ , versione modificata della statistica radice con segno del log-rapporto di verosimiglianza  $r$ . L'obiettivo che si persegue è di confrontare gli intervalli di confidenza basati su  $r$  e  $r^*$  per l'indice di Youden, che rappresenta il parametro di interesse.

L'indice di Youden è un indice molto utilizzato in campo scientifico, soprattutto epidemiologico. Esso rappresenta, insieme all'AUC (area sotto curva ROC), una misura globale e sintetica della bontà diagnostica di un test. In un test diagnostico la scelta di un valore di soglia individua sensibilità, o probabilità dei veri positivi, e specificità, o probabilità dei veri negativi. L'indice di Youden in relazione al valore di soglia ottimale corrisponde alla ricerca del massimo della somma di sensibilità e specificità, ossia al miglior test possibile.

La tesi si articola in quattro capitoli principali.

Il capitolo 1 porta l'attenzione nell'ambito dei test diagnostici e presenta brevemente gli strumenti utilizzati per valutarne l'accuratezza. Si introducono quindi gli indici sintetici di sensibilità e specificità e la curva ROC. Si procede con la descrizione dell'indice di Youden e la presentazione di alcune procedure, presenti in letteratura, riguardanti la sua stima.

Il capitolo 2 richiama i concetti base dell'inferenza di verosimiglianza. Si introduce la nozione di verosimiglianza profilo per l'eliminazione di parametri di disturbo. Si presentano le statistiche test classiche, quali il log-rapporto di verosimiglianza  $W$  e la sua versione unilaterale  $r$ .

Il capitolo 3 descrive brevemente la teoria asintotica di ordine superiore, in particolare la formula  $p^*$  di Barndorff-Nielsen, che consente di costruire una modificazione di  $r$ , chiamata radice con segno modificata del log-rapporto di verosimiglianza,  $r^*$ . Si presenta l'approssimazione di Skovgaard per il calcolo di  $r^*$  e il pacchetto `likelihoodAsy` di R che implementa quest'ultima.

Il capitolo 4 è dedicato agli studi di simulazione effettuati per verificare le coperture empiriche degli intervalli di confidenza basati sulle statistiche  $r$  e  $r^*$ . Viene utilizzato il pacchetto `LikelihoodAsy` per effettuare lo studio in diverse situazioni: nel caso di popolazioni che seguono distribuzioni normali o gamma per differenti valori dei parametri e per diverse numerosità campionarie.

# Capitolo 1

## Test diagnostici e indice di Youden

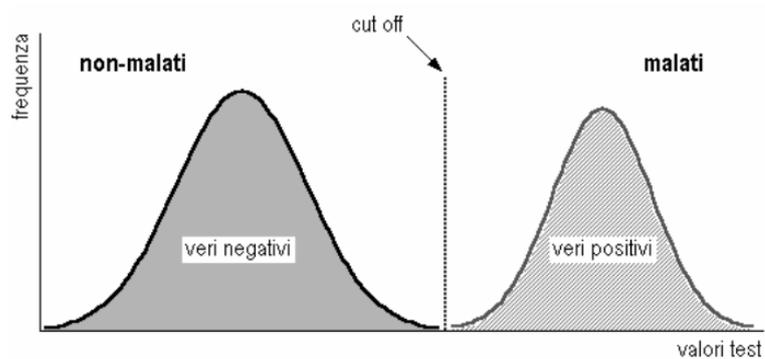
### 1.1 Test diagnostici

In ambito scientifico vengono continuamente introdotte procedure più o meno complesse e di svariata natura per verificare, dimostrare o confermare determinate ipotesi. Tali metodi sono comunemente detti *test*. In campo medico, in particolare epidemiologico, si sviluppano test diagnostici o di *screening* per studiare l'incidenza di una determinata malattia su una popolazione: si cerca di individuare la presenza della malattia nei soggetti sottoposti a esame, cioè di discriminare i malati dai sani.

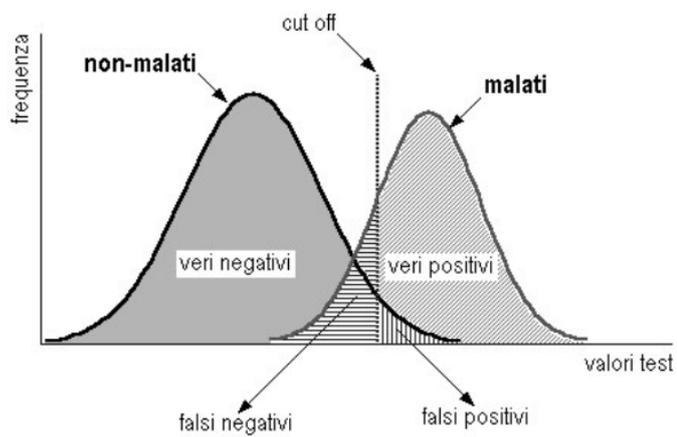
Si definisce quindi test diagnostico una procedura il cui scopo è di classificare un individuo come sano o malato; in base alla tipologia di risposta fornita, essi possono essere distinti in due categorie:

- **qualitativo** se il risultato sul singolo individuo è di natura binaria (positivo/negativo, affetto/non affetto);
- **quantitativo** se restituisce una misura numerica, sia essa di natura discreta o continua.

Nel caso in cui l'esito sia una variabile discreta o continua si definisce un valore di soglia, detto *cut-off* o *cut point*, per confermare o escludere la presenza sui soggetti osservati della malattia sospettata.



**Figura 1.1:** Situazione ideale - *cut-off* immediatamente determinato.



**Figura 1.2:** Situazione reale - ricerca del *cut-off* ottimale.

Il test diagnostico ideale (detto *golden standard*) sarebbe in grado di dividere perfettamente malati e non malati attraverso il valore di *cut-off*, poiché individui positivi e negativi darebbero valori del test differenti (Figura 1.1). Nella realtà invece è quasi impossibile che le due popolazioni siano completamente separate e le distribuzioni degli esiti del test sono parzialmente sovrapposte (Figura 1.2). Si hanno pertanto persone sane che risultano positive al test, dette "falsi positivi", e persone malate che risultano negative al test, dette "falsi negativi".

Il risultato di un test diagnostico si può riassumere in una tabella di contingenza (Tabella 1.1), dove le colonne rappresentano i soggetti sani e malati e le righe indicano il risultato del test sui pazienti. Le quattro tipologie di individui osservati sono quindi: malati positivi o veri positivi ( $V_+$ ), sani positivi o falsi positivi ( $F_+$ ), malati negativi o falsi negativi ( $F_-$ ), sani negativi veri negativi ( $V_-$ ).

	Malati	Sani
Positivi	$V_+$	$F_+$
Negativi	$F_-$	$V_-$

**Tabella 1.1:** Tabella di contingenza.

## 1.2 Sensibilità, specificità e curva ROC

Si presenta allora la necessità di calcolare il grado di incertezza della classificazione, ossia la bontà del test ad un determinato valore di soglia; per misurarla si introducono indici sintetici, quali la sensibilità e la specificità.

La **sensibilità** è la probabilità che un soggetto realmente malato risulti positivo al test (probabilità dei "veri positivi" o "*true positive*"). È misurata dal rapporto tra veri positivi e il totale degli individui effettivamente malati, indicato con  $TM_+$ .

$$\begin{aligned}
 S &= \frac{V_+}{(V_+ + F_-)} \\
 &= \frac{V_+}{TM_+}.
 \end{aligned}$$

Un test sarà tanto più sensibile quanto più bassa risulterà la quota dei falsi negativi (cioè di soggetti malati erroneamente identificati dal test come sani). Un test molto sensibile in definitiva ci consente di limitare la possibilità che un soggetto malato risulti negativo al test.

La **specificità** è la probabilità che un soggetto realmente sano risulti negativo al test (probabilità dei "veri negativi" o "*true negative*"). È misurata dal rapporto tra veri negativi e il totale degli individui effettivamente sani, indicato con  $TS_+$ .

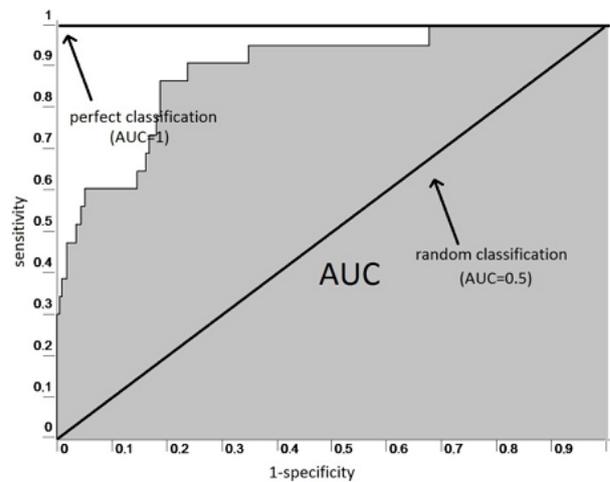
$$\begin{aligned} S &= \frac{V_-}{(V_- + F_+)} \\ &= \frac{V_-}{TS_+}. \end{aligned}$$

Sensibilità e specificità sono fra loro inversamente correlate in rapporto alla scelta del valore di *cut-off*. Infatti, cambiando il *cut-off* si può ottenere un aumento della sensibilità e diminuzione della specificità, così come una diminuzione della sensibilità ed aumento della specificità.

Per analizzare il rapporto tra "veri positivi" e "falsi positivi" si introduce un metodo grafico o **curva ROC** (dall'acronimo inglese *Receiver Operating Characteristics*).

La curva ROC rappresenta la capacità di un test a condurre una diagnosi corretta, che è di identificare i "veri positivi" e i "veri negativi" al variare dei valori di soglia. In un sistema di riferimento cartesiano, sugli assi vengono riportati in ordinata e in ascissa rispettivamente la sensibilità e (1-specificità), ossia *True Positive Rate* (veri positivi) e *False Positive Rate* (falsi positivi). Si studia quindi la funzione che mette in relazione la sensibilità al complemento a uno della specificità.

Qualsiasi regola di decisione statistica propone, per una ripartizione ideale degli individui, la massimizzazione di sensibilità e specificità in quanto si ipotizza una situazione in cui non vi siano falsi né negativi, né positivi. Nella realtà ciò non è possibile e poiché sensibilità e specificità sono fra loro inversamente correlate, si cerca di individuare un compromesso tra i due indici. Si possono individuare graficamente due curve limite interpretabili (Figura 1.3):



**Figura 1.3:** Curve limite e area sottesa alla curva ROC.

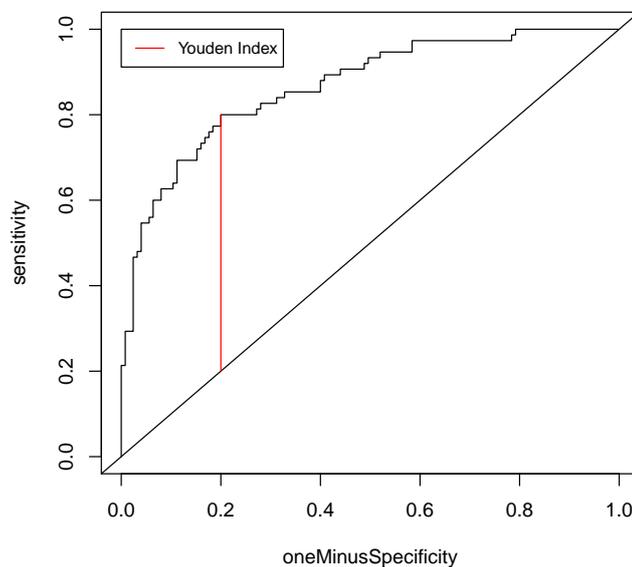
- Una prima che taglia il grafico a  $45^\circ$ , passando per l'origine; questa retta rappresenta il caso del test casuale (linea di "nessun beneficio"), e l'area sottesa è pari a 0,5;
- Una seconda curva è rappresentata dall'insieme di segmenti che dall'origine sale al punto (0,1) e da quello che congiunge il punto (0,1) a (1,1), avendo un'area sottesa di valore pari a 1, ovvero rappresenta il test perfetto.

Per maggiori approfondimenti, si rimanda all'articolo di Bottarelli e Parodi (2003).

### 1.3 Indice di Youden

Si introducono due indici sintetici comunemente usati per riassumere l'informazione relativa alla curva ROC: l'area sottesa alla curva stessa e l'indice di Youden, di cui ci si occuperà in seguito in questa relazione.

L'area sottesa alla curva ROC (**AUC** per *Area Under Curve*), fornisce un'informazione utile per verificare la capacità discriminativa del test: maggiore è il valore dell'area, maggiore risulta l'accuratezza del test (Figura 1.3).



**Figura 1.4:** Indice di Youden.

L'**indice di Youden** ( $J$ ), introdotto da W.J. Youden nel 1950 in ambito medico, rappresenta la massimizzazione di una funzione della sensibilità  $q(c)$  e della specificità  $p(c)$  rispetto ad un valore di soglia  $c$ , che di conseguenza è quel *cut-point* che identifica più correttamente la distinzione tra sani e malati.

$$\begin{aligned}
 J &= \max_c \{ \text{sensibilità}(c) + \text{specificità}(c) - 1 \} \\
 &= \max_c \{ q(c) + p(c) - 1 \} \\
 &= \max_c \{ q(c) - (1 - p(c)) \}
 \end{aligned}$$

$J$  assume valori compresi nell'intervallo  $[0,1]$ . Il valore 0 indica che il test diagnostico restituisce la stessa proporzione di risultati positivi per i gruppi affetti e non affetti da malattia e corrisponde quindi ad un test inefficace. Il valore 1 indica che non vi sono falsi positivi o falsi negativi, ossia il test è perfettamente efficace. L'indice dà lo stesso peso a falsi positivi e falsi negativi, ossia tutte le prove con lo stesso valore dell'indice corrispondono alla stessa proporzione di individui risultati erroneamente classificati.

L'indice di Youden viene usato in combinazione con la curva ROC, poiché è definito per tutti i punti della curva. Esso si calcola empiricamente sommando specificità e sensibilità meno uno per tutti i possibili *cut-points*. Graficamente,  $J$  rappresenta la massima distanza verticale tra la curva ROC e la diagonale o *chance line*, come evidenziato dalla Figura 1.4.

A differenza dell'AUC l'indice di Youden permette di calcolare direttamente il valore di *cut-off*. Il valore massimo dell'indice può essere utilizzato come criterio per la selezione del punto di soglia o *cut-off* ottimale.

## 1.4 Stima dell'indice di Youden

Siano  $x_1, \dots, x_m$  e  $y_1, \dots, y_n$  i campioni osservati per le variabili casuali continue  $X$  e  $Y$  indipendenti corrispondenti al risultato di un test diagnostico, rispettivamente per gli ammalati e i sani. Siano  $F$  e  $G$  le rispettive funzioni di ripartizione di  $X$  e  $Y$ . Per un determinato valore di soglia  $c$ , l'indice di Youden si può riscrivere:

$$\begin{aligned} J &= \max_c \{ \text{sensibilità}(c) + \text{specificità}(c) - 1 \} \\ &= \max_c \{ P(X \geq c) + P(Y \leq c) - 1 \}. \end{aligned}$$

Se  $c_0$  è il valore del *cut-off* che realizza questo massimo, allora  $J$  diventa:

$$\begin{aligned} J &= \{ \text{sensibilità}(c_0) + \text{specificità}(c_0) - 1 \} \\ &= \{ P(X \geq c_0) + P(Y \leq c_0) - 1 \}. \end{aligned}$$

Per stimare l'indice di Youden sono possibili due approcci. L'**approccio non parametrico**, senza ipotesi distributive su  $X$  e  $Y$ , utilizza le funzioni di ripartizioni empiriche di  $X$  e  $Y$ .

Per un *cut-point* stimato  $\hat{c}_0$ , le stime per la sensibilità e la specificità si scrivono

$$\text{sensibilità} = \frac{\sum_{i=1}^m I(x_i \geq \hat{c}_0)}{m}, \quad \text{specificità} = \frac{\sum_{j=1}^n I(y_j \leq \hat{c}_0)}{n},$$

e per l'indice di Youden

$$\begin{aligned} \hat{J} &= \frac{\sum_{i=1}^m I(x_i \geq \hat{c}_0)}{m} + \frac{\sum_{j=1}^n I(y_j \leq \hat{c}_0)}{n} - 1 \\ &= \frac{\sum_{j=1}^n I(y_j \leq \hat{c}_0)}{n} - \frac{\sum_{i=1}^m I(x_i \leq \hat{c}_0)}{m}. \end{aligned}$$

L'**approccio parametrico**, con funzioni di ripartizione  $F$  e  $G$  note, porta ad una stima della sensibilità e della specificità del tipo

$$\text{sensibilità} = 1 - F(\hat{c}_0), \quad \text{specificità} = G(\hat{c}_0),$$

e dell'indice di Youden del tipo

$$\hat{J} = G(\hat{c}_0) - F(\hat{c}_0).$$

Per maggiori approfondimenti, si rimanda all'articolo di Zhou (2008).

Per quanto riguarda la costruzione di intervalli di confidenza sull'indice di Youden in ambito parametrico, Schisterman e Perkins (2007) ne hanno effettuato lo studio nei casi in cui le distribuzioni di  $X$  e  $Y$  seguano distribuzioni normali o gamma. In questo contesto si collocheranno le simulazioni effettuate nel Capitolo 4 della tesi. Essi hanno realizzato inoltre intervalli di confidenza utilizzando il metodo Delta per stimare la varianza di  $\hat{J}$ . In alternativa, in ambito non parametrico, hanno introdotto intervalli di confidenza basati su due versioni bootstrap, il metodo *Basic Percentiles* (BP) e il metodo *Bias Corrected and Accelerated* (BCa).

# Capitolo 2

## L'inferenza di verosimiglianza

### 2.1 Introduzione

Quando si conduce un esperimento o uno studio osservazionale, è spesso complicato o dispendioso effettuare l'analisi su tutta la popolazione. L'inferenza statistica è l'insieme dei metodi che ci permettono, a partire da un'osservazione parziale del fenomeno di interesse, di generalizzarne i risultati: dalle caratteristiche del campione osservato si risale a quelle della popolazione sotto indagine da cui il campione proviene. Questa procedura è tuttavia soggetta ad un'evidente imprecisione, poichè le informazioni contenute nel campione casuale non sempre riassumono in modo completo quelle di partenza. Diventa necessario definire uno schema generale, applicabile sotto determinate premesse e che preveda una componente erratica.

In questo capitolo si tratterà dell'inferenza basata sulla verosimiglianza, introdotta e formulata esplicitamente da Ronald Fisher (1890-1962). Si daranno quindi le definizioni di modello statistico e di funzione di verosimiglianza, per poi affrontare il caso dell'inferenza in presenza di parametri di disturbo e presentare la verosimiglianza profilo e la proprietà di invarianza. Quindi si introdurranno stime puntuali, intervalli di confidenza e verifiche d'ipotesi globali e parziali. Tutti questi strumenti rientrano nella teoria della verosimiglianza basata sulle approssimazioni del primo ordine.

Per una trattazione più completa degli argomenti discussi in questo capitolo si rimanda alla lettura del testo di Pace e Salvani (2001, cap. 3-5-6).

## 2.2 Specificazione del modello

L'idea che sta alla base dell'inferenza statistica è che il vettore di osservazioni  $y = (y_1, \dots, y_n)$ , ottenuto ripetendo un esperimento casuale nelle stesse condizioni un numero finito  $n$  di volte, è la realizzazione di una variabile casuale  $Y$ . La distribuzione teorica di  $Y$ , detta distribuzione di probabilità e indicata con  $p_0(y)$ , è ignota.

L'obiettivo dell'inferenza statistica è di estrarre dai dati informazioni sul modello probabilistico  $p_0(y)$  che li ha generati. Occorre quindi specificare un **modello statistico**  $\mathcal{F}$ , ovvero una famiglia di distribuzioni di probabilità compatibili con  $Y$  che contenga il vero modello probabilistico  $p_0(y)$ . In particolare, se  $p_0(y) \in \mathcal{F}$  allora si dice che il modello statistico è **correttamente specificato**.

Assumiamo che il modello  $\mathcal{F}$  utilizzato in seguito sia un **modello parametrico**, ossia che i suoi elementi si possano descrivere con un numero finito di parametri. Un modello statistico parametrico può essere dunque espresso come

$$\mathcal{F} = \{p(y, \theta) : \theta \in \Theta \subseteq \mathbb{R}^p\}.$$

L'insieme  $\Theta$  di tutti i possibili valori che il parametro  $\theta$  può assumere è detto spazio parametrico. Se la dimensione  $p$  del parametro è 1, allora il parametro è detto scalare, se  $p > 1$  allora  $\theta = (\theta_1, \dots, \theta_p)$  è un vettore di parametri. Il modello è **identificabile** se esiste una relazione biunivoca tra ciascun elemento di  $\mathcal{F}$  e il valore assunto dal parametro.

## 2.3 Verosimiglianza

### 2.3.1 Funzione di verosimiglianza

Una volta osservato il campione  $y = (y_1, \dots, y_n)$  e definito un modello parametrico  $\mathcal{F}$ , che rispetta le ipotesi di corretta specificazione e identificabilità, la funzione di densità  $p(y; \theta)$  dipende solo dal parametro  $\theta$ .

Si definisce **funzione di verosimiglianza** di  $\theta$ , relativamente ad un  $y$  fissato, una funzione  $L : \Theta \rightarrow \mathbb{R}^+$  tale che

$$L(\theta) = L(\theta; y) = c(y)p(y; \theta).$$

Il sostegno empirico a favore di  $\theta_1$  rispetto a  $\theta_2$  è misurato dal rapporto  $L(\theta_1)/L(\theta_2)$ , detto **rapporto di verosimiglianza**. Alla luce delle osservazioni,  $\theta_1 \in \Theta$  è più plausibile di  $\theta_2 \in \Theta$  se  $L(\theta_1)/L(\theta_2) > 1$ .

Per questo motivo la costante moltiplicativa positiva  $c(y)$  si può quindi trascurare e in particolare verosimiglianze proporzionali sono dette **verosimiglianze equivalenti**. La notazione  $L(\theta; y)$  serve ad evidenziare che, pur avendo la stessa forma di una distribuzione di probabilità, nella funzione di verosimiglianza  $y$  è fissato e  $\theta$  varia.

Se si assume che le osservazioni  $y = (y_1, \dots, y_n)$  siano realizzazioni indipendenti ed identicamente distribuite (i.i.d) allora la funzione di verosimiglianza ha la forma

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

dove  $p_{Y_i}(y_i; \theta)$  è la distribuzione di probabilità marginale per la singola osservazione.

In genere, a fini pratici, viene utilizzata la trasformazione logaritmica di  $L(\theta; y)$ , detta funzione di **log-verosimiglianza**,

$$l(\theta) = l(\theta; y) = \log L(\theta; y).$$

Anche in questo caso è possibile trascurare la costante positiva  $c(y)$  e considerare la **log-verosimiglianza equivalente**, che differisce dalla funzione di log-verosimiglianza per una costante additiva

$$l(\theta; y) = c'(y) + \log p(y; \theta)$$

con  $c'(y) = \log c(y)$  e con  $l(\theta) = -\infty$  se  $p(y; \theta) = 0$ . L'interpretazione di  $l(\theta; y)$  è ora in termini di differenze: a partire dalle osservazioni,  $\theta_1 \in \Theta$  è più plausibile di  $\theta_2 \in \Theta$  se  $l(\theta_1) - l(\theta_2) > 0$ . Nel caso il campione sia costituito da osservazioni i.i.d, per la proprietà dei logaritmi si ottiene la log-verosimiglianza nella seguente forma

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

### 2.3.2 Stima di massima verosimiglianza

A questo punto si può procedere alla stima di  $\theta$ , che individua un unico modello probabilistico all'interno di  $\mathcal{F}$ : si cerca il valore  $\hat{\theta}$  più verosimile, ossia quello più supportato dai dati osservati.

Si ricorda che si è di fronte ad un **problema di stima regolare** se valgono le seguenti condizioni di regolarità (si veda, ad esempio, Azzalini 2008, p. 88):

- il modello è identificabile;
- lo spazio parametrico  $\Theta$  è un sottoinsieme aperto dello spazio euclideo  $\mathbb{R}^p$ , con  $p$  finito e indipendente da  $n$ ;
- le funzioni di probabilità (o densità) specificate in  $\mathcal{F}$  hanno tutte lo stesso supporto, che non dipende da  $\theta$ ;
- la log-verosimiglianza è derivabile almeno tre volte con derivate parziali continue in  $\Theta$  e sono leciti scambi per la funzione di densità  $p(y; \theta)$  fino a due volte del segno di integrale rispetto a  $y$  con quello di derivata rispetto a  $\theta$ .

Dato un modello con verosimiglianza regolare e funzione di verosimiglianza  $L(\theta)$ , si chiama **stima di massima verosimiglianza** di  $\theta$  (SMV) il valore  $\hat{\theta} \in \Theta$  tale che

$$L(\hat{\theta}) \geq L(\theta), \quad \forall \theta \in \Theta.$$

ovvero quel valore di  $\theta$  che massimizza la funzione di verosimiglianza  $L(\theta)$ . La stima di massima verosimiglianza può essere determinata anche a partire dalla log-verosimiglianza, di cui pure costituisce un punto di massimo. Non è detto che  $\hat{\theta}$  esista o che sia unica e per alcuni modelli può essere difficile ottenere una soluzione esplicita per cui si deve ricorrere a delle procedure di calcolo numeriche per riuscire a massimizzare  $l(\theta)$ .

Sempre se le condizioni di regolarità sono soddisfatte, è possibile definire delle **quantità di verosimiglianza**, che contengono le principali informazioni sull'andamento della funzione di verosimiglianza.

- la **funzione punteggio**: è il vettore delle derivate parziali di primo ordine della funzione  $l(\theta)$

$$l_*(\theta) = \left( \frac{\partial l(\theta)}{\partial \theta_1}, \dots, \frac{\partial l(\theta)}{\partial \theta_p} \right) = \left[ \frac{\partial l(\theta)}{\partial \theta_r} \right], = [l_r(\theta)]$$

con  $1 < r < p$ . Uguagliando a zero la funzione punteggio si ottengono le **equazioni di verosimiglianza** la cui soluzione unica rispetto a  $\theta$  individua la stima di massima verosimiglianza

$$l_*(\theta) = 0;$$

- la **matrice di informazione osservata**  $j(\theta)$ : è la matrice delle derivate parziali seconde di  $l(\theta)$  cambiate di segno.

$$j(\theta) = -l_{**}(\theta) = \left[ -\frac{\partial^2 l(\theta)}{\partial \theta_r \partial \theta_s} \right] = [-l_{rs}(\theta)]$$

dove  $-l_{rs}(\theta)$  è l'elemento di posizione  $(r, s)$  contenente l'opposto delle derivate seconde parziali di  $l(\theta)$ ;

- la **matrice di informazione attesa** o informazione di Fisher  $i(\theta)$ : è il valore atteso della matrice di informazione osservata.

$$i(\theta) = E_\theta\{j(\theta)\} = \left[ -E_\theta \left( \frac{\partial^2 l(\theta)}{\partial \theta_r \partial \theta_s} \right) \right] = [i_{rs}(\theta)]$$

dove  $i_{rs}(\theta)$  è l'elemento di posizione  $(r, s)$  di  $i(\theta)$ .

### 2.3.3 Statistiche sufficienti

Si introduce ora uno strumento di estrazione di informazione sul vero valore del parametro dal campione di osservazioni, detto statistica sufficiente.

Dato un modello statistico parametrico  $\mathcal{F} = \{p(y, \theta) : \theta \in \Theta \subseteq \mathbb{R}^p\}$ , una funzione è detta **statistica** se è del tipo  $T: \mathcal{Y} \rightarrow \mathbb{R}^r$  per un qualche  $r$  naturale, ossia se dipende dal solo campione osservato e non da  $\theta$ .  $T(y)$  è quindi una trasformazione che si spera sintetizzi i dati osservati e rappresenta una partizione dello spazio campionario. La condizione di indipendenza di  $T$  da  $\theta$  garantisce l'effettiva computabilità della statistica. Il suo valore  $t = T(y)$  è detto **valore campionario**.

Per poter riassumere i dati, senza però perdere delle informazioni sul parametro di interesse  $\theta$ , si utilizzano le cosiddette statistiche sufficienti. Una statistica  $s$  è detta **sufficiente** per  $\theta$  se la distribuzione di  $Y$  condizionata a  $S = s$  non dipende dal parametro  $\theta$  e pertanto vale la fattorizzazione

$$p_Y(y; \theta) = p_s(s; \theta)p_{Y|S=s}(y; s)$$

per tutti gli  $y$  tali che  $S(y) = s$ .

Si noti che  $p_{Y|S=s}(y; s)$  non dipende dal parametro e si può quindi trascurare in quanto semplice costante moltiplicativa. La statistica sufficiente per il parametro  $\theta$  assume quindi lo stesso valore in punti dello spazio campionario che hanno verosimiglianze equivalenti; per ogni  $y_1, y_2 \in \mathcal{Y}$

$$S(y_1) = S(y_2) \rightarrow L(\theta; y_1) \propto L(\theta; y_2).$$

Ciò conduce al **principio di sufficienza**: se  $s$  è statistica sufficiente, due osservazioni  $y_1$  e  $y_2$  per cui  $S(y_1) = S(y_2)$  devono condurre alle medesime conclusioni inferenziali su  $\theta$ .

Un modello statistico  $\mathcal{F}$  ammette di norma svariate statistiche sufficienti e tra tutte sono preferibili le statistiche sufficienti più concise, poiché più adeguate a condensare tutta l'informazione portata dai dati sul parametro ignoto. A tal proposito si definisce statistica **sufficiente minimale** per il modello  $\mathcal{F}$  quella statistica  $S$  che è funzione di ogni altra statistica sufficiente e rappresenta perciò la massima sintesi dei dati senza perdita di informazione sul parametro che ci permette di fare inferenza su di esso.

Essa assume valori diversi per due punti dello spazio campionario solo quando le rispettive verosimiglianze non sono equivalenti. In notazione, per ogni  $y_1, y_2 \in \mathcal{Y}$

$$S(y_1) = S(y_2) \iff L(\theta; y_1) \propto L(\theta; y_2).$$

Un'altra classe di statistiche è quella delle statistiche ancillari. Una statistica  $a$  è detta **ancillare** rispetto a  $S$ , sufficiente minimale, se la relazione tra  $(\hat{\theta}, a)$  e  $S$  è biunivoca e se la distribuzione di  $a$  non dipende da  $\theta$ , cioè  $a$  è costante in distribuzione. Se  $a$  è ancillare allora vale la fattorizzazione

$$p_{\hat{\theta}, A}(\hat{\theta}, a; \theta) = p_A(a)p_{\hat{\theta}|A=a}(\hat{\theta}; a, \theta),$$

dove  $\hat{\theta}$  diventa quindi statistica sufficiente minimale nel modello condizionato al valore osservato di  $a$ . Tale quantità recupera l'informazione riguardante la precisione con cui la stima di massima verosimiglianza individua  $\theta_0$  entro  $\Theta$ , perduta nel passaggio da  $s$  a  $\hat{\theta}$ .

## 2.4 Parametri di disturbo

Come si è visto nei paragrafi precedenti, per rappresentare una realtà di interesse viene in aiuto un modello statistico  $\mathcal{F}$ , cioè una famiglia di modelli probabilistici del tipo  $p(y; \theta)$  dove  $y$  rappresenta i dati osservati e  $\theta$  un vettore di parametri che definisce il modello. La dimensione del parametro  $\theta$  dipende dal grado di complessità della realtà sotto studio: se questa complessità cresce, per rendere il modello adeguato a descrivere i dati osservati, aumenta anche la dimensione del parametro  $\theta$ .

Tuttavia spesso l'analisi non è rivolta a studiare tutti gli aspetti della distribuzione proposta, quindi tutti parametri, ma solamente una parte. Diventa allora opportuno distinguere il parametro  $\theta$  in due componenti: i parametri d'interesse, che rappresentano gli aspetti che saranno oggetto di inferenza, e i parametri di disturbo, che servono solamente a descrivere meglio la variabilità della popolazione. Si considererà quindi in seguito una partizione del parametro  $p$ -dimensionale  $\theta = (\psi, \lambda)$ , dove  $\psi$  rappresenta il **parametro d'interesse** di dimensione  $k$  e  $\lambda$  è il **parametro di disturbo** di dimensione  $p - k$ .

Poichè l'inferenza riguarda solo il parametro  $\psi$  sorge la necessità di formulare una nuova verosimiglianza che prenda in considerazione solamente il parametro di interesse. Solitamente si cerca di ridurre il problema ad una funzione di verosimiglianza dipendente solo da  $\psi$ , cercando di evitare o limitare le perdite sul parametro di interesse. Una funzione dipendente, oltre che dai dati, dal solo parametro di interesse e che abbia una o più proprietà che la rendano simile ad una verosimiglianza in senso proprio è detta **pseudo-verosimiglianza**.

## 2.5 Verosimiglianza profilo

Uno degli approcci classici per la costruzione di una pseudo verosimiglianza per il parametro d'interesse  $\psi$  propone di sostituire il parametro di disturbo  $\lambda$  con una sua stima consistente. Si sostituisce quindi  $\lambda$  con  $\hat{\lambda}_\psi$  ottenuto massimizzando la log-verosimiglianza  $l(\psi, \lambda)$  in  $\lambda$  per  $\psi$  fissato. La funzione che ne deriva, detta **funzione di verosimiglianza profilo**, si scrive

$$L_p(\psi) = L_p(\psi, \hat{\lambda}_\psi).$$

La verosimiglianza profilo consente di fare inferenza su  $\psi$  e di ottenerne la stima di massima verosimiglianza  $\hat{\theta} = (\hat{\psi}, \hat{\lambda}) = (\hat{\psi}, \hat{\lambda}_{\hat{\psi}})$ . Alla stessa soluzione si può arrivare con la funzione di log-verosimiglianza profilo  $l_p(\psi) = \log L_p(\psi)$ .

La verosimiglianza profilo non può essere considerata come un verosimiglianza in senso proprio poichè non viene calcolata partendo da una funzione di densità. Tuttavia gode di proprietà che la rendono simile ad una verosimiglianza propria:

- la **stima di massima verosimiglianza profilo** coincide con la stima di massima verosimiglianza di  $\psi$  basata su  $L(\psi, \theta)$ , ovvero

$$\sup_{\psi} L_p(\psi) = L_p(\hat{\psi});$$

- si può definire l'**informazione osservata profilo** come

$$j_P(\psi) = -\frac{\partial^2}{\partial \psi \partial \psi^T} l_P(\psi) = -\frac{\partial^2}{\partial \psi \partial \psi^T} l(\psi, \hat{\lambda}_\psi).$$

La matrice inversa di  $j_P(\psi)$  è uguale al blocco  $(\psi, \psi)$  dell'inversa della matrice di informazione osservata complessiva. Si osserva infatti che

$$\frac{\partial}{\partial \psi} l_P(\psi) = l(\psi, \hat{\lambda}_\psi) + l_\lambda(\psi, \hat{\lambda}_\psi) \frac{\partial}{\partial \psi} \hat{\lambda}_\psi,$$

dove il secondo addendo è uguale a zero poichè  $l_\lambda(\psi, \hat{\lambda}_\psi) = 0$ . Risulta pertanto

$$\frac{\partial}{\partial \psi} l_P(\psi) = l_\psi(\psi, \hat{\lambda}_\psi).$$

La matrice delle derivate seconde è data da

$$\frac{\partial^2}{\partial \psi} l_P(\psi) = l_{\psi\psi}(\psi, \hat{\lambda}_\psi) + l_{\psi\lambda}(\psi, \hat{\lambda}_\psi) \frac{\partial}{\partial \psi} \hat{\lambda}_\psi$$

e dato che  $l_\lambda(\psi, \hat{\lambda}_\psi) = 0$ , anche la sua derivata calcolata rispetto a  $\psi$  sarà pari a 0. Si ottiene così

$$l_{\psi\lambda}(\psi, \hat{\lambda}_\psi) + l_{\lambda\lambda}(\psi, \hat{\lambda}_\psi) \frac{\partial}{\partial \psi} \hat{\lambda}_\psi = 0$$

da cui

$$\frac{\partial}{\partial \psi} \hat{\lambda}_\psi = -(l_{\lambda\lambda}(\psi, \hat{\lambda}_\psi))^{-1} l_{\psi\lambda}(\psi, \hat{\lambda}_\psi)$$

di conseguenza

$$j_P(\psi) = -(l_{\psi\psi} - l_{\psi\lambda}(l_{\lambda\lambda})^{-1}l_{\psi\lambda})$$

dove tutte le derivate si intendono calcolate in  $(\psi, \hat{\lambda}_\psi)$ . Si ottiene quindi

$$[j_P(\psi)]^{-1} = j^{\psi\psi}(\psi, \hat{\lambda}_\psi),$$

dove  $j^{\psi\psi}(\psi, \hat{\lambda}_\psi)$  rappresenta il blocco  $(\psi, \psi)$  dell'inversa di  $j(\psi, \hat{\lambda}_\psi)$ .

Quando  $\psi$  è uno scalare, si ha che la curvatura della log-verosimiglianza profilo è direttamente legata alla precisione di  $\hat{\psi}$ . Una proprietà della log-verosimiglianza  $l(\theta)$ , quando non ci sono parametri di disturbo, è che l'informazione osservata  $j(\hat{\theta})$  può essere usata come una stima dell'inversa della matrice di covarianza asintotica di  $\hat{\theta}$ , che corrisponde a  $i(\theta)$  (Young e Smith 2005, p. 136). Il risultato di sopra mostra che la funzione corrispondente calcolata a partire dalla log-verosimiglianza profilo

$$[j_P(\hat{\psi})] = -[\partial_\psi \partial_\psi^T l_P(\psi)]_{\psi=\hat{\psi}}$$

determina una stima dell'inversa della matrice di covarianza asintotica per  $\hat{\psi}$ .

- il test **log-rapporto di verosimiglianza profilo** è

$$W_P(\psi) = 2\{l_P(\hat{\psi}) - l_P(\psi)\} = 2\{l(\hat{\psi}, \hat{\lambda}) - l(\psi, \hat{\lambda}_\psi)\}$$

ovvero coincide con l'usuale log-rapporto di verosimiglianza basato su  $L(\psi, \lambda)$  considerato per la verifica d'ipotesi per  $\psi$ , con  $\lambda$  ignoto. Sotto condizioni di regolarità  $W_P(\psi)$  ha distribuzione asintotica nulla pari ad un  $\chi_k^2$ , dove  $k$  è la dimensione del parametro di interesse.

Le proprietà precedenti, avvalendosi di una approssimazione del primo ordine, mostrano perchè la verosimiglianza profilo risulta essere uno strumento utile per l'inferenza su  $\psi$ . La verosimiglianza profilo permette infatti di effettuare test o costruire intervalli di confidenza su  $\psi$  basati su un'approssimazione  $\chi^2$  della distribuzione della statistica  $W_P(\psi)$  oppure su un'approssimazione normale nel caso di appropriate funzioni pivotali approssimate che verranno introdotte nel Paragrafo 2.6.

Tali approssimazioni possono tuttavia condurre a risultati insoddisfacenti, specialmente se la dimensione del parametro  $\lambda$  è elevata. In effetti  $l_p(\psi)$  non è in generale il logaritmo di una densità e utilizzare la verosimiglianza profilo equivale a considerare  $\lambda$  noto e pari a  $\hat{\lambda}$ , trascurando la variabilità della stima di  $\lambda$ . Esistono diverse possibilità di modificazione della verosimiglianza profilo. Una di queste, che si introduce poichè sarà utile nel Capitolo 3, è la **verosimiglianza profilo modificata** proposta da Barndorff-Nielsen (1980, 1983) e che differisce dalla log-verosimiglianza profilo per un fattore di correzione. Essa è definita come:

$$L_M(\psi) = L_P(\psi)M(\psi),$$

dove  $M(\psi)$  è un fattore di aggiustamento pari a

$$M(\psi) = |l_{\lambda;\hat{\lambda}}(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda}, a)|^{-1} |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda}, a)|^{1/2},$$

dove nel primo determinante del fattore di modificazione compare la derivata mista

$$l_{\lambda;\hat{\lambda}}(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda}, a) = \frac{\partial^2}{\partial \lambda \partial \hat{\lambda}^T} l(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda}, a)$$

per cui è richiesta l'esplicitazione di una statistica ancillare  $a$ , esatta o approssimata. Le quantità  $l_{\lambda;\hat{\lambda}}$  che compaiono nell'espressione coinvolgono derivate rispetto alle componenti dello stimatore di massima verosimiglianza  $\hat{\theta}$ , che sono dette **derivate nello spazio campionario**.

## 2.6 Invarianza rispetto a riparametrizzazioni

In un modello statistico parametrico identificabile esiste una relazione biunivoca tra i suoi elementi,  $p(y; \theta)$ , e il parametro  $\theta$ . Scegliere una parametrizzazione significa un sistema di coordinate geometriche all'interno del quale

gli elementi di  $\mathcal{F}$  possano sempre essere individuati univocamente. In alcuni casi si presenta la necessità di riparametrizzare il modello per evitare vincoli sul parametro rendendo lo spazio parametrico illimitato, ma è desiderabile che la riparametrizzazione non cambi le conclusioni inferenziali.

Una trasformazione biunivoca  $\omega = \omega(\theta)$ , con  $\omega: \Theta \rightarrow \Omega$  e  $\Theta, \Omega \in \mathbb{R}^p$ , è una riparametrizzazione del modello. Si parla di **invarianza rispetto alla parametrizzazione** quando le conclusioni inferenziali su  $\omega$  sono le stesse di quelle su  $\theta$  e possono essere espresse in funzione di  $\theta = \theta(\omega)$ .

Questa proprietà, se riferita all'ambito di stima puntuale prodotta da una certa procedura inferenziale, viene detta di **equivarianza**. Ne deriva che se  $\hat{\theta}$  è la stima di massima verosimiglianza di  $\theta$  allora  $\hat{\omega} = \omega(\hat{\theta})$  sarà la stima di massima verosimiglianza di  $\omega$  e  $L_{\Omega}(\omega(\hat{\theta})) = L_{\Theta}(\hat{\theta})$ , dove  $L_{\Omega}(\omega(\theta))$  e  $L_{\Theta}(\theta)$  sono rispettivamente la funzione di verosimiglianza riscritta nella nuova parametrizzazione e la funzione di verosimiglianza nella parametrizzazione originale.

In situazioni in cui sono presenti parametri di disturbo non è detto che si possa riparametrizzare il modello senza cambiarne le conclusioni inferenziali. Generalmente una riparametrizzazione globale non mantiene la distinzione tra parametro di interesse e di disturbo, quindi bisogna restringere il campo delle riparametrizzazioni che non alterano l'interesse, ossia le riparametrizzazioni  $\omega = \omega(\theta) = \omega(\psi, \lambda)$ , con  $w = (\rho, \xi)$  tali che  $\rho = \rho(\psi)$  e  $\xi = \xi(\psi, \lambda)$ , con  $\rho(\cdot)$  funzione biunivoca.

## 2.7 Inferenza di verosimiglianza

Per risalire alla vera ignota distribuzione  $p(y, \theta)$  una buona soluzione è sostituire la stima di massima verosimiglianza  $\hat{\theta}$  nella funzione di densità del modello  $\mathcal{F}$ . La teoria asintotica di verosimiglianza, infatti, garantisce nei problemi regolari di stima che lo stimatore  $\hat{\theta}$  sia **consistente**: per  $n \rightarrow +\infty$   $\hat{\theta}$  converge in probabilità a  $\theta$ , vero valore del parametro. Inoltre, per  $n$  sufficientemente grande, la distribuzione di  $\hat{\theta}$  per si può approssimare come segue:

$$\hat{\theta} \sim N_p(\theta, i(\theta)^{-1})$$

dove  $p$  indica la dimensione del parametro. In tale espressione,  $i(\theta)$  può essere sostituita dalla sua stima  $i(\hat{\theta})$ , da  $j(\theta)$  o  $j(\hat{\theta})$ .

Il metodo della massima verosimiglianza permette, oltre a trovare una stima puntuale per il parametro di interesse, di effettuare **verifiche d'ipotesi** o di costruire **intervalli di confidenza** sul parametro.

### 2.7.1 Test d'ipotesi

Il test d'ipotesi consiste nel verificare la plausibilità di un certo valore  $\theta_0$  per il parametro ignoto  $\theta$ . Le due decisioni possibili relative al parametro  $\theta$  si possono sintetizzare in un **sistema d'ipotesi**, formulato prima di osservare i dati:

$$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta \neq \theta_0. \end{cases}$$

La proposizione  $H_0$  viene detta **ipotesi nulla**, mentre la proposizione  $H_1$  è chiamata **ipotesi alternativa** e può essere bilaterale o unilaterale. Ciò che permette di stabilire se è più ragionevole l'ipotesi nulla o quella alternativa è la statistica test, che consente di dividere lo spazio campionario in due sottoinsiemi disgiunti. Si definisce **statistica test** una funzione  $T: \mathcal{Y} \rightarrow R$ , dipendente sia dai dati  $y$  che dal parametro  $\theta$ , ma con distribuzione sotto  $H_0$  nota e indipendente dal parametro, detta **distribuzione nulla**.

I **valori critici** della distribuzione nulla individuano due regioni, la **regione di accettazione** e la **regione di rifiuto** dell'ipotesi nulla. Se il valore campionario  $T(y)$  appartiene alla regione di accettazione allora l'ipotesi nulla è confermata dai dati e si accetta l'ipotesi; viceversa, se  $T(y)$  appartiene alla regione di rifiuto allora l'ipotesi alternativa è sostenuta dall'evidenza empirica e si conclude che il test è significativo contro  $H_0$ .

Il test statistico non è una procedura infallibile, perché può essere che il campione sorteggiato cada nella regione di accettazione A o di rifiuto R per effetto del caso. Si commette un **errore di I tipo** se si rifiuta  $H_0$  quando questa è vera e un **errore di II tipo** se si accetta  $H_0$  quando questa è falsa.

La massima probabilità di commettere un errore del I tipo è chiamata **livello di significatività**, e si indica con

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta[T(y) \in R].$$

La verosimiglianza può essere usata per la costruzione di statistiche test approssimate. Per testare l'ipotesi semplice  $H_0 : \theta = \theta_0$  contro l'alternativa bilaterale  $H_1 : \theta \neq \theta_0$  si può definire un test legato alla verosimiglianza, il test **log-rapporto di verosimiglianza**. Esso è una trasformazione monotona del rapporto di verosimiglianza  $L(\hat{\theta})/L(\theta_0)$ ; si ottiene come log-verosimiglianza normalizzata, calcolata per  $\theta = \theta_0$  e moltiplicata per -2

$$W(\theta_0) = 2\{l(\hat{\theta}) - l(\theta_0)\}.$$

Il test  $W(\theta_0)$  si distribuisce asintoticamente, sotto condizioni di regolarità, come un chi-quadrato con  $p$  gradi di libertà, dove  $p$  è il numero di componenti di  $\theta$ . La regione di rifiuto è unilaterale destra e si rifiuta l'ipotesi nulla  $H_0$  per valori grandi di  $W(\theta_0)$ . La statistica  $W(\theta)$  è, per ogni  $\theta \in \Theta$  fissato, una quantità asintoticamente pivotale.

Quando  $p = 1$  e l'ipotesi alternativa è unilaterale (destra o sinistra) si può utilizzare la statistica test **radice con segno del log-rapporto di verosimiglianza**

$$r(\theta_0) = \text{sgn}(\hat{\theta} - \theta_0) \sqrt{W(\theta_0)},$$

dove  $\text{sgn}(\cdot)$  è la funzione segno, con  $\text{sgn}(x) = +1$  se  $x > 0$ ,  $\text{sgn}(x) = -1$  se  $x < 0$ ,  $\text{sgn}(x) = 0$  se  $x = 0$ . Il test  $r(\theta_0)$  ha distribuzione nulla approssimata asintoticamente, sotto condizioni di regolarità, a una  $N(0, 1)$ . La regione critica è unilaterale destra o sinistra a seconda dell'alternativa  $H_1$  e si rifiuta l'ipotesi nulla per valori estremi di  $r(\theta_0)$ .

Esistono anche delle versioni asintoticamente equivalenti del test, come il **test di Wald** e il **test punteggio** (*score test*) o **test di Rao**

$$W_e(\theta_0) = (\hat{\theta} - \theta_0)^T i(\theta_0)^{-1} (\hat{\theta} - \theta_0),$$

$$W_u(\theta_0) = l_*(\theta_0)^T i(\theta_0)^{-1} l_*(\theta_0),$$

entrambi con distribuzione nulla asintotica chi-quadrato con  $p$  gradi di libertà. Per  $p = 1$  si possono definire le versioni di Wald e punteggio di  $r(\theta_0)$  di

forma

$$\begin{aligned} r_e(\theta_0) &= (\hat{\theta} - \theta_0)\sqrt{i(\theta_0)}, \\ r_u(\theta_0) &= l_*(\theta_0)\sqrt{i(\theta_0)}, \end{aligned}$$

con distribuzione asintotica nulla  $N(0, 1)$ .

In caso di parametri di disturbo, grazie alle proprietà della verosimiglianza profilo che la assimilano ad una verosimiglianza propria, si possono costruire dei test asintoticamente equivalenti a quelli costruiti con il parametro globale. Per verificare l'ipotesi composita  $H_0 : \psi = \psi_0$  contro l'alternativa bilaterale  $H_1 : \psi \neq \psi_0$  si può definire il test **log-rapporto di verosimiglianza profilo**

$$W_P(\psi_0) = 2\{l_P(\hat{\psi}) - l_P(\psi_0)\} = 2\{l(\hat{\psi}, \hat{\lambda}) - l(\psi_0, \hat{\lambda}_{\psi_0})\}.$$

Il test  $W(\theta_0)$  si distribuisce asintoticamente, sotto condizioni di regolarità, come un chi-quadrato con  $k$  gradi di libertà, dove  $k$  è il numero di componenti di  $\psi$ . La statistica  $W_P(\psi)$  è, per ogni  $\psi \in \Psi$  fissato, una quantità asintoticamente pivotale.

Se il parametro d'interesse  $\psi$  è scalare, si può considerare anche la radice con segno di  $W_P(\psi_0)$

$$r_P(\psi_0) = \text{sgn}(\hat{\psi} - \psi_0)\sqrt{W_P(\psi_0)},$$

che ha distribuzione nulla approssimata asintoticamente, sotto condizioni di regolarità, a una  $N(0, 1)$ .

Analogamente si possono calcolare le statistiche di Rao e di Wald

$$W_{Pe}(\psi_0) = (\hat{\psi} - \psi_0)^T [i^{\psi\psi}(\psi_0, \hat{\lambda}_{\psi_0})]^{-1} (\hat{\psi} - \psi_0),$$

$$W_{Pu}(\psi_0) = l_\psi(\psi_0, \hat{\lambda}_{\psi_0})^T i^{\psi\psi}(\psi_0, \hat{\lambda}_{\psi_0})^{-1} l_\psi(\theta_0, \hat{\lambda}_{\psi_0}),$$

entrambi con distribuzione nulla asintotica chi-quadrato con  $k$  gradi di libertà e dove per  $i^{\psi\psi}$  si intende il blocco relativo a  $\psi$  di  $i(\theta)^{-1}$ . Per  $p = 1$  si possono definire le versioni di Wald e punteggio di forma:

$$r_{Pe}(\psi_0) = (\hat{\psi} - \psi_0) [i^{\psi\psi}(\psi_0, \hat{\lambda}_{\psi_0})]^{-1/2},$$

$$r_{Pu}(\psi_0) = l_\psi(\psi_0, \hat{\lambda}_{\psi_0}) \sqrt{i^{\psi\psi}(\psi_0, \hat{\lambda}_{\psi_0})},$$

con distribuzione asintotica nulla  $N(0, 1)$ .

I test rapporto di verosimiglianza, di Wald e di Rao sono asintoticamente equivalenti. Si preferiscono tuttavia i test basati su  $W_P(\psi)$  o  $r_P(\psi)$  per la loro proprietà di invarianza alla riparametrizzazione, piuttosto che le versioni di Wald e di Rao. Infatti il test di Wald non è invariante rispetto a riparametrazioni e il test di Rao è a volte instabile numericamente.

### 2.7.2 Stima intervallare

La procedura di stima intervallare è invece volta a determinare una **regione di confidenza** per il vero valore del parametro, ovvero un sottoinsieme di  $\Theta$  individuato sulla base dei dati  $y$ , di valori plausibili per  $\theta$ . Esso si può rappresentare come

$$\hat{\Theta}(y) \subset \Theta.$$

Se il parametro è scalare, si parla di **intervallo di confidenza**.

Anche in questo caso, viene scelto il livello di confidenza nominale  $(1 - \alpha)$  della regione, tale che

$$\mathbb{P}_\theta(\theta \in \hat{\Theta}(Y)) = 1 - \alpha \quad \forall \theta \in \Theta.$$

Per la costruzione di intervalli di confidenza si utilizzano usualmente quantità pivotali  $q(y; \theta)$ . Si definisce **quantità pivotale** una funzione dipendente sia dai dati  $y$  che dal parametro  $\theta$ , con distribuzione nota indipendente dal parametro stesso.

Le regioni di confidenza per  $\theta$  con livello nominale  $(1 - \alpha)$  possono essere calcolate a partire dalla quantità pivotale  $W(\theta)$ , visto che la distribuzione asintotica di  $W(\theta_0)$  è  $\chi_p^2$ , e sono di forma:

$$\hat{\theta}(y) = \{\theta \in \Theta : W(\theta) \leq \chi_{p;1-\alpha}^2\}.$$

La regione può anche essere riscritta come

$$\hat{\theta}(y) = \{\theta \in \Theta : l(\theta) \geq l(\hat{\theta}) - \frac{1}{2}\chi_{p;1-\alpha}^2\},$$

dove  $\chi_{p;1-\alpha}^2$  è il quantile  $(1 - \alpha)$  della distribuzione  $\chi_p^2$ .

Nel caso in cui siano presenti parametri di disturbo l'obiettivo è di individuare una regione di confidenza per il parametro di interesse  $\psi$ . La statistica

$W_P(\psi_0)$  è una quantità pivotale per  $\psi$  e ciò permette di costruire una regione di confidenza per  $\psi$  con livello nominale  $(1 - \alpha)$ , di forma

$$\hat{\theta}(y) = \{\psi \in \Theta_\psi : W_P(\psi) \leq \chi_{k;1-\alpha}^2\}.$$

dove  $\chi_{k;1-\alpha}^2$  è il quantile  $(1 - \alpha)$  della distribuzione  $\chi_k^2$ . Se  $\psi$  è scalare si ha allora l'intervallo di confidenza

$$\hat{\theta}(y) = \{\psi \in \Theta_\psi : |r_P(\psi)| \leq z_{1-\frac{\alpha}{2}}\},$$

dove  $z_{1-\frac{\alpha}{2}}$  è il quantile di livello  $(1 - \frac{\alpha}{2})$  della distribuzione  $N(0, 1)$ .

# Capitolo 3

## Teoria asintotica di ordine superiore

### 3.1 Introduzione

Nel capitolo precedente si è visto che è difficile riuscire ad ottenere la distribuzione esatta delle statistiche che derivano dalla verosimiglianza, per cui si studia il loro comportamento quando la numerosità campionaria è molto alta. Per modelli statistici parametrici regolari, il teorema del limite centrale e la legge dei grandi numeri permettono di ottenere una serie di risultati asintotici che riguardano la stima di massima verosimiglianza, intervalli di confidenza e test d'ipotesi. Tali approssimazioni risultano tuttavia poco accurate in certe situazioni e a partire dagli anni '80 dello scorso secolo sono emerse nuove procedure basate su approssimazioni di ordine superiore. Lo scopo è di rendere più accurate le usuali approssimazioni normale o chi-quadrato per le statistiche di tipo  $r$  e  $W$ , nel caso di numerosità campionaria piccola o moderata, e di tenere conto della presenza di parametri di disturbo.

In questo capitolo verranno presentati i principali risultati partendo dalla formula  $p^*$  per arrivare ad ottenere la statistica radice con segno modificata  $r^*$  in presenza di parametri di disturbo. Verrà in seguito esplicitata l'approssimazione di Skovgaard che consente il calcolo di  $r^*$  senza richiedere l'esplicitazione di statistiche ancillari. Infatti, le derivate nello spazio campionario vengono approssimate da opportuni valori attesi. Questi ultimi, a loro

volta, vengono approssimati tramite l'uso dei metodi di integrazione Monte Carlo. La sua implementazione è disponibile nel pacchetto `LikelihoodAsy` di R, realizzato da Bellio e Pierce (2015a).

I test di riferimento per una presentazione più dettagliata degli argomenti di questo capitolo sono quelli di Brazzale, Davison et al. (2007, cap. 2-8) e Young e Smith (2005, cap. 8-9).

### 3.2 La formula $p^*$

Il risultato principale che è alla base della teoria asintotica di ordine superiore è la **formula**  $p^*$  di Barndorff-Nielsen. Questa formula consiste in un'approssimazione per la densità dello stimatore di massima verosimiglianza condizionata ad una statistica ancillare.

Un concetto fondamentale per capire ciò che segue è che lo stimatore di massima verosimiglianza  $\hat{\theta}$  è sufficiente se e solo se  $i(\hat{\theta}) = j(\hat{\theta})$ ; questo accade solamente nella famiglia esponenziale nel caso di una parametrizzazione di rango pieno. Al di fuori della famiglia esponenziale è possibile, condizionandosi ad una statistica ancillare del tipo  $a = j(\hat{\theta})/i(\hat{\theta})$  (ossia  $i(\hat{\theta})^{-1}j(\hat{\theta})$ ), rendere lo stimatore di massima verosimiglianza sufficiente oltre il primo ordine. L'approssimazione proposta da Barndorff-Nielsen si colloca in questo secondo caso.

Sia  $s = s(y)$  una statistica sufficiente minimale per  $\theta$  e  $a = a(y)$  una statistica ancillare. Esprimendo  $s$  in funzione di  $(\hat{\theta}; a)$  si riscrive  $l(\theta; s) = l(\theta; \hat{\theta}, a)$ . Barndorff-Nielsen ha ottenuto la seguente approssimazione, chiamata  $p^*$ , per la densità condizionata  $f(\hat{\theta}|a; \theta)$  dello stimatore di massima verosimiglianza  $\hat{\theta}$ :

$$p^*(\hat{\theta}|a; \theta) = c(\theta, a) |j(\hat{\theta})|^{1/2} \exp\{l(\theta) - l(\hat{\theta})\}, \quad (3.1)$$

dove  $c(\theta, a)$  è una costante di normalizzazione che rende pari a 1 l'integrale della (3.1) rispetto a  $\hat{\theta}$ , mantenendo costanti  $a$  e  $\theta$ .

Questa equazione dà la distribuzione condizionata esatta dello stimatore di massima verosimiglianza per un'ampia classe di modelli (Young e Smith 2005, p. 155). Più generalmente la formula dà un'approssimazione con errore

relativo di ordine  $O(n^{-1})$  e, per ogni  $\hat{\theta}$  fissato,

$$f(\hat{\theta}|a; \theta) = p^*(\hat{\theta}|a; \theta)\{1 + O(n^{-1})\}.$$

L'aspetto notevole dell'approssimazione  $p^*$  è che la formula precedente coincide con l'espressione ottenuta nel caso in cui  $\hat{\theta}$  è sufficiente:

$$\begin{aligned} p^*(\hat{\theta}; \theta) &= c(\theta)|j(\hat{\theta})|^{1/2} \exp\{l(\theta) - l(\hat{\theta})\} \\ &= p^*(\hat{\theta}; \theta)\{1 + O(n^{-1})\}. \end{aligned}$$

### 3.3 La statistica radice con segno modificata $r^*$

Si considera in seguito il caso in cui  $\theta = (\psi, \lambda)$ , con  $\psi$  parametro di interesse scalare e  $\lambda$  parametro di disturbo. Per effettuare una verifica di ipotesi sul parametro di interesse  $\psi$ , con ipotesi del tipo  $H_0 : \psi = \psi_0$  e ipotesi alternativa unidirezionale, si può utilizzare la statistica  $r$  radice con segno del log-rapporto di verosimiglianza

$$r_P(\psi_0) = \text{sgn}(\hat{\psi} - \psi_0) \sqrt{W(\psi_0)},$$

che si distribuisce asintoticamente come una normale standard con errore pari a  $O(n^{-1/2})$ . In particolare

$$Pr\{r_P(\psi_0) < r_P(\psi_0)^{oss}\} = \Phi(r_P(\psi_0)^{oss}) + O_P(n^{-1/2}),$$

dove  $\Phi(\cdot)$  è la funzione di ripartizione di una  $\mathcal{N}(0, 1)$  e  $r_P(\psi_0)^{oss}$  è il valore osservato di  $r_P(\psi_0)$ .

La formula  $p^*$  consente di ottenere una versione modificata per la statistica radice con segno, introdotta anch'essa da Barndorff-Nielsen e chiamata **radice con segno modificata del rapporto di log-verosimiglianza**  $r^*$ , che si scrive nella forma

$$r^*(\psi) = r(\psi) + \frac{1}{r(\psi)} \log \left\{ \frac{u(\psi)}{r(\psi)} \right\},$$

o, attraverso un'ulteriore scomposizione

$$r^*(\psi) = r(\psi) + \frac{1}{r(\psi)} \log(C(\psi)) + \frac{1}{r(\psi)} \log \left\{ \frac{\tilde{u}(\psi)}{r(\psi)} \right\},$$

dove

$$C(\psi) = \left| \frac{\partial^2 l(\psi, \hat{\lambda}_\psi)}{\partial \lambda \partial \hat{\lambda}^T} \right| \{ |\hat{j}_{\lambda\lambda}| |\tilde{j}_{\lambda\lambda}| \}^{-1/2},$$

$$\tilde{u}(\psi) = \left| \frac{\partial}{\partial \hat{\psi}} \{ l_P(\psi) - l_P(\hat{\psi}) \} \right| |\hat{j}_{\psi\psi|\lambda}|^{-1/2}.$$

I primi fattori delle due espressioni coinvolgono delle derivazioni rispetto ai dati (*sample space derivatives*).

Le matrici  $\tilde{j}_{\lambda\lambda}$  e  $\hat{j}_{\lambda\lambda}$  sono rispettivamente le matrici di informazione osservata per  $\lambda$  sotto stima di massima verosimiglianza vincolata  $\hat{\theta} = (\hat{\psi}, \hat{\lambda})$  e non vincolata  $\hat{\theta}_\psi = \tilde{\theta} = (\psi, \hat{\lambda}_\psi)$ . La matrice  $\hat{j}_{\psi\psi|\lambda}$  è l'informazione osservata aggiustata per  $\psi$ :  $\hat{j}_{\psi\psi|\lambda} = \hat{j}_{\psi\psi} - \hat{j}_{\psi\lambda} \hat{j}_{\lambda\lambda}^{-1} \hat{j}_{\lambda\psi}$ .

La statistica  $r^*$  ottenuta ha ancora distribuzione asintotica normale standard ma con un errore di ordine  $O_P(n^{-3/2})$ . In particolare

$$Pr[r_P(\psi_0) < r_P(\psi_0)^{oss}] = \Phi(r_P^*(\psi_0)^{oss}) + O_P(n^{-3/2}).$$

L'approssimazione è quindi più accurata poichè l'errore tende a zero più velocemente al crescere di  $n$ .

Si può riscrivere  $r^*(\psi)$  sotto la forma

$$r^*(\psi) = r(\psi) + NP(\psi) + INF(\psi),$$

dove  $NP$  rappresenta un termine di **correzione per i parametri di disturbo** e  $INF$  un termine di **correzione per la precisione**, ossia un aggiustamento per migliorare l'approssimazione normale.

Il penultimo termine  $NP$  proviene dalla verosimiglianza profilo modificata di Barndorff-Nielsen  $L_{MP}(\psi; y) \propto L_P(\psi; y) \exp(-C_\psi)$  ed è rilevante quando la dimensione del parametro di disturbo è elevata, anche per campioni abbastanza grandi. L'aggiustamento  $INF$  è invece spesso di minor importanza purchè l'informazione osservata  $\hat{j}_P(\hat{\psi})$  non sia troppo piccola relativamente alla dimensione di  $\lambda$ . Entrambe i termini della decomposizione sono concettualmente importanti, ma in pratica il termine  $INF$  ha minore effetto del termine  $NP$ .

### 3.4 Approssimazione di Skovgaard

L'aspetto più difficile per ottenere  $r^*$  risiede nella specificazione di  $C(\psi)$  e  $\tilde{u}(\psi)$ , che risulta complessa perché richiede un calcolo delle derivate rispetto allo spazio campionario. Il calcolo esatto è possibile solo nel caso di famiglie esponenziali, poiché questi termini fanno intervenire solo quantità legate alla verosimiglianza. Per la dimostrazione nel caso esponenziale si rinvia a Brazzale, Davison et al. (2007).

Nel caso generale la determinazione risulta difficile e sono state proposte varie approssimazioni. In particolare Skovgaard (1996) ha suggerito di approssimare le derivate rispetto allo spazio campionario con apposite covarianze tra log-verosimiglianze e funzioni punteggio. L'espressione di queste covarianze è la seguente:

$$\frac{\partial^2 l(\psi, \hat{\lambda}_\psi)}{\partial \theta \partial \hat{\theta}^T} \doteq \text{cov}_{\theta_2}(l_\theta(\theta_1), l_\theta(\theta_2))|_{\theta_1=\tilde{\theta}, \theta_2=\hat{\theta}} \hat{\tau}^{-1} \hat{J}, \quad (3.2)$$

$$\frac{\partial l(\hat{\psi}, \hat{\lambda})}{\partial \hat{\theta}} - \frac{\partial l(\psi, \hat{\lambda}_\psi)}{\partial \hat{\theta}} \doteq \text{cov}_{\theta_2}(l(\theta_2) - l(\theta_1), l_\theta(\theta_1))_{\theta_1=\tilde{\theta}, \theta_2=\hat{\theta}} \hat{\tau}^{-1} \hat{J}, \quad (3.3)$$

dove  $l_\theta(\theta) = \partial l(\theta; y) / \partial \theta$ . Nelle espressioni (3.2) e (3.3) prima vengono calcolate le covarianze con  $\theta_1$  e  $\theta_2$  generici e poi il risultato è valutato in  $\theta_1 = \tilde{\theta}$  e  $\theta_2 = \hat{\theta}$ . L'approssimazione di Skovgaard evita quindi il calcolo delle derivate rispetto ai dati e il condizionamento alla statistica ancillare.

Ottenere queste covarianze richiede tuttavia il calcolo di alcuni valori attesi non standard, che possono però essere approssimati mediante un metodo di **simulazione Monte Carlo**.

Riscrivendo le covarianze di (3.2) e (3.3) con le notazioni

$$S(\theta_1, \theta_2) = \text{cov}_{\theta_2}\{l_\theta(\theta_1), l_\theta(\theta_2)\},$$

$$Q(\theta_1, \theta_2) = \text{cov}_{\theta_2}\{l_\theta(\theta_1), l(\theta_2) - l(\theta_1)\},$$

si ottengono:

$$S(\theta_1, \theta_2) = E_{\theta_2}[l_\theta(\theta_1)l_\theta(\theta_2)] = \int l_\theta(\theta_1)l_\theta(\theta_2) p(y; \theta_1) dy, \quad (3.4)$$

$$Q(\theta_1, \theta_2) = E_{\theta_2}[l_\theta(\theta_1)(l(\theta_2) - l(\theta_1))] = \int l_\theta(\theta_1)[l(\theta_2) - l(\theta_1)] p(y; \theta_1) dy. \quad (3.5)$$

La determinazione delle covarianze empiriche avviene sui campioni simulati con il metodo Monte Carlo. Tuttavia le stime di massima verosimiglianza  $\hat{\theta}$  e  $\tilde{\theta}$  rimangono le stesse, ossia quelle ottenute sul campione originale  $y$  osservato. Il metodo differisce quindi dal bootstrap parametrico, che calcolerebbe ad ogni simulazione una nuova verosimiglianza con nuove stime  $\hat{\theta}$  e  $\tilde{\theta}$  diverse ad ogni passo.

### 3.5 Pacchetto **LikelihoodAsy**

Il pacchetto `LikelihoodAsy` di R, sviluppato da Bellio e Pierce (2015c) contiene funzioni utili per determinare la verosimiglianza profilo modificata e le statistiche  $r$  e  $r^*$ . La funzione `rstar` consente di realizzare test d'ipotesi con i relativi  $p$ -value per un parametro scalare in presenza di parametri di disturbo.

Il requisito per la sua esecuzione è che l'utente fornisca tre argomenti:

- la funzione di log-verosimiglianza del modello assunto (`floglik`);
- la funzione che genera i dati a partire dal modello (`datagen`);
- la funzione che definisce il parametro di interesse su cui si vuole fare inferenza (`fpsi`).

Il pacchetto utilizza l'approssimazione di Skovgaard e l'integrazione Monte Carlo per il calcolo delle covarianze (3.4) e (3.5); il secondo argomento è la funzione che genera i dati per il calcolo Monte Carlo di (3.4) e (3.5).

L'output della funzione `rstar`, che verrà utilizzata nel Capitolo 4, è il seguente:

- le stime di  $\hat{\theta}$ ,  $\tilde{\theta}$  e  $\hat{\psi}$ ;
- i valori delle statistiche di Wald,  $r$  e  $r^*$  con i relativi  $p$ -value corrispondenti al valore sotto  $H_0$ ;
- i termini di correzione  $NP$  e  $INF$ .

---

La funzione `rstar.ci` invece fornisce intervalli di confidenza per  $\psi$  basati su  $r$  e  $r^*$ . Entrambe le funzioni permettono di fare inferenza su una generica funzione  $\psi(\theta)$  (argomento `fpsi`); ossia, non è necessario che  $\psi$  sia esplicitamente una componente di  $\theta$ .



# Capitolo 4

## Studi di simulazione

### 4.1 Introduzione

In questo capitolo si intende valutare, da un punto di vista empirico, il comportamento delle stime dell'indice di Youden mediante uno studio di simulazione. Si effettuerà l'analisi per due distribuzioni che possono assumere i malati e i sani: normale e gamma.

Preso nota del fatto che le tecniche di inferenza di verosimiglianza di primo ordine peggiorano in affidabilità nel caso di bassa numerosità campionaria, l'obiettivo centrale dello studio sarà proprio quello di valutare l'accuratezza degli strumenti di inferenza di verosimiglianza di ordine superiore in relazione a diverse numerosità campionarie. Si confronteranno quindi le coperture empiriche degli intervalli di confidenza costruiti mediante le statistiche  $r_P$  e  $r_P^*$ , ossia la percentuale di intervalli che contiene il vero valore del parametro. Si valuterà inoltre l'approssimazione alla normale delle due statistiche. Risulteranno più precisi gli intervalli di confidenza con copertura empirica che si avvicina maggiormente ai valori nominali.

Si rimanda all'Appendice per avere visione del codice R sviluppato per effettuare le simulazioni, A.1 ed A.2 per il caso normale, A.3 per il caso gamma.

## 4.2 Procedura di confronto fra $r$ e $r^*$

Come visto nel Capitolo 3,  $r_P^*$  è una quantità pivotale, avente distribuzione nulla normale standard con errore di ordine  $O(n^{-3/2})$ . Un intervallo di confidenza per  $\psi$  basato su  $r_P^*(\psi)$  con un livello approssimato  $(1 - \alpha)$  è quindi dato da  $(\psi_L^*, \psi_U^*)$ , soluzioni in  $\psi$  delle equazioni

$$r_P^*(\psi) = z_{1-\frac{\alpha}{2}}, \quad r_P^*(\psi) = z_{\frac{\alpha}{2}}.$$

In modo equivalente, un intervallo di confidenza ha la forma

$$\{\psi: |r_P^*(\psi_0)| \leq z_{1-\frac{\alpha}{2}}\}$$

La procedura sviluppata in questo capitolo per valutare le coperture empiriche degli intervalli di confidenza basati sulle le statistiche  $r_P$  e  $r_P^*$ , utilizza la funzione `rstar` all'interno di un ciclo. Si procede nel seguente modo:

- per  $r = 1, \dots, R$  volte si genera il campione  $(y_1^r, \dots, y_n^r)$  da  $f(y; \psi_0, \lambda_0)$ ;
- si esegue il test per il valore fissato  $\psi = \psi_0$  e si ottengono i valori delle statistiche  $r_P$  e  $r_P^*$  calcolate sul campione replicato.

Si calcolano quindi le percentuali di copertura empirica, ossia le medie dei valori di  $r_P$  e  $r_P^*$  compresi nell'intervallo di confidenza di livello nominale  $(1 - \alpha)$  di una  $N(0, 1)$  prefissato

$$1 - \hat{\alpha} = \frac{1}{R} \sum_{r=1}^R I\{|r_P(\psi_0)| \leq z_{1-\frac{\alpha}{2}}\} = \frac{1}{R} \sum_{r=1}^R I(\hat{\psi}_L^r < \psi_0 < \hat{\psi}_U^r),$$

$$1 - \hat{\alpha}^* = \frac{1}{R} \sum_{r=1}^R I\{|r_P^*(\psi_0)| \leq z_{1-\frac{\alpha}{2}}\} = \frac{1}{R} \sum_{r=1}^R I(\hat{\psi}_L^{*r} < \psi_0 < \hat{\psi}_U^{*r}).$$

Per valutare la qualità degli intervalli di confidenza basati sulle statistiche  $r_P$  e  $r_P^*$  si confrontano le percentuali di copertura empiriche ottenute. Inoltre, si riportano anche gli errori di copertura sinistro e destro.

Si calcola inoltre la variabilità della copertura empirica (indicata nelle Tabelle con `std`), misurata attraverso lo standard error della stima  $\hat{\beta}$  della copertura  $1 - \alpha$ , poichè essa ha distribuzione approssimativamente normale.

### 4.3 Distribuzione normale

Assumiamo che  $X$  e  $Y$ , che rappresentano rispettivamente gli ammalati e i sani, siano variabili casuali indipendenti e distribuiti normalmente secondo una  $N(\mu_X, \sigma_X^2)$  e  $N(\mu_Y, \sigma_Y^2)$ . Si considera il caso in cui la media di  $X$  è maggiore della media di  $Y$  ( $\mu_X > \mu_Y$ ); nel caso contrario è sufficiente invertire le notazioni  $X$  e  $Y$ . Si generano due campioni di dati  $x_1, \dots, x_m$  proveniente da  $X$  e  $y_1, \dots, y_n$  proveniente da  $Y$ .

Nel caso in cui  $\sigma_X^2 \neq \sigma_Y^2$  il cut-point si scrive (Schisterman, Perkins et al. 2005, Appendice 1)

$$c_{1,2} = \frac{\mu_Y(b^2 - 1) - a + b\sqrt{a^2 + (b^2 - 1)\sigma_Y^2 \ln(b^2)}}{(b^2 - 1)},$$

con  $a = \mu_X - \mu_Y$  e  $b = \sigma_X/\sigma_Y$ .

Nel caso in cui  $\sigma^2 = \sigma_X^2 = \sigma_Y^2$  si ha  $b = 1$  e il *cut-point*  $c$  corrisponde al punto medio delle due medie  $\mu_X$  e  $\mu_Y$ :

$$c = \frac{\mu_X + \mu_Y}{2}.$$

L'indice di Youden corrispondente si può esprimere come

$$\psi = \psi(\theta) = \Phi\left(\frac{\mu_X - c}{\sigma_X}\right) + \Phi\left(\frac{c - \mu_Y}{\sigma_Y}\right) - 1.$$

con  $\Phi(\cdot)$  funzione di ripartizione di una normale standard e  $\theta = (\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2)$ .

La stima di massima verosimiglianza di  $\psi$  è data da

$$\hat{\psi} = \psi(\hat{\theta}) = \Phi\left(\frac{\hat{\mu}_X - c}{\hat{\sigma}_X}\right) + \Phi\left(\frac{c - \hat{\mu}_Y}{\hat{\sigma}_Y}\right) - 1,$$

dove  $\hat{\mu}_X = (\sum_i^m x_i)/m$ ,  $\hat{\mu}_Y = (\sum_j^n x_j)/n$  sono le stime di massima verosimiglianza di  $\mu_X$ ,  $\mu_Y$  e  $\hat{\sigma}_X^2 = (\sum_i^m (x_i - \hat{\mu}_X)^2)/m$ ,  $\hat{\sigma}_Y^2 = (\sum_j^n (x_j - \hat{\mu}_Y)^2)/n$  sono le stime di massima verosimiglianza di  $\sigma_X^2$  e  $\sigma_Y^2$ .

La riparametrizzazione nel caso in cui  $\sigma_X^2 \neq \sigma_Y^2$  si scrive  $\theta = (\psi, \lambda)$ , dove  $\psi$  corrisponde all'indice di Youden e  $\lambda = (\mu_Y, \sigma_X^2, \sigma_Y^2)$ . La riparametrizzazione nel caso in cui  $\sigma_X^2 = \sigma_Y^2$  si scrive  $\theta = (\psi, \lambda)$ , dove  $\psi$  corrisponde all'indice di Youden e  $\lambda = (\mu_Y, \sigma^2)$ , con  $\sigma^2$  varianza comune alle due popolazioni.

Nella seguente sezione si eseguiranno le simulazioni per studiare i due casi normali:  $X$  e  $Y$  con varianze uguali e  $X$  e  $Y$  con varianze diverse. Si assumerà che  $Y$  si distribuisca come una  $N(0, 1)$ , quindi con  $\mu_Y = 0$  e  $\sigma_Y^2 = 1$ . I valori delle medie  $\mu_X$  verranno determinati in modo da ottenere un indice di Youden fissato pari a  $J = (0.4, 0.6, 0.8)$  e i valori delle varianze verranno scelte a  $\sigma_X = (0.5, 1, 3)$ , come indicato nella Tabella 4.1.

Youden Index	Standard deviation di X		
	$\sigma_X^2 = 0.5$	$\sigma_X^2 = 1$	$\sigma_X^2 = 3$
$J = 0.4$	0.8484	1.0489	1.2540
$J = 0.6$	1.4071	1.6833	2.1843
$J = 0.8$	2.1682	2.5632	3.4247

**Tabella 4.1:** Medie  $\mu_X$  per varianze  $\sigma_X^2$  e indici  $J$  di Youden fissati.

Le numerosità campionarie per  $X$  e  $Y$ , rispettivamente  $(m, n)$ , verranno fissate a  $(50, 50)$ ,  $(30, 30)$ ,  $(20, 20)$ ,  $(10, 10)$ ,  $(5, 5)$ . I livelli nominali di riferimento sono scelti al 90%, 95% e 99%.

Il numero di replicazioni è fissato a  $R = 5000$ . Per l'algoritmo Monte Carlo che calcola le covarianze empiriche il numero di ripetizioni è mantenuto al valore di default  $B = 1000$ .

### 4.3.1 Normali con varianze uguali

Il primo studio di simulazione svolto è quello in cui  $X$  e  $Y$  hanno varianze uguali e  $\psi = (0.4, 0.6, 0.8)$ . I risultati sono riportati nelle Tabelle da 4.2 a 4.4.

La Tabella 4.2 si riferisce al caso in cui  $\psi = 0.4$ . È immediato notare che gli intervalli basati su  $r_P^*$  raggiungono i livelli di copertura desiderati per ogni livello di confidenza anche per campioni a bassa numerosità, dove invece quelli basati su  $r_P$  perdono in precisione. Le coperture empiriche mostrano infatti che gli intervalli basati su  $r$  sono buoni per numerosità pari o superiori a  $(m, n) = (30, 30)$ , mentre degradano per numerosità inferiori.

(m,n)	Metodo	Livello di copertura nominale		
		90%	95%	99%
(5,5)	$r_P$	87.32 (0.00, 12.68)	92.10 (0.00, 7.90)	97.44 (0.00, 2.56)
	$r_P^*$	93.57 (0.82, 5.61)	96.75 (0.17, 3.08)	99.16 (0.17, 0.67)
(10,10)	$r_P$	88.79 (2.81, 8.40)	94.63 (0.71, 4.67)	98.79 (0.00, 1.21)
	$r_P^*$	91.20 (3.98, 4.83)	96.28 (1.23, 2.48)	99.21 (0.08, 0.71)
(20,20)	$r_P$	88.94 (3.60, 7.46)	94.08 (1.90, 4.02)	98.36 (0.32, 1.32)
	$r_P^*$	89.82 (4.94, 5.24)	94.46 (2.56, 2.98)	98.42 (0.54, 1.04)
(30,30)	$r_P$	89.58 (3.96, 6.46)	94.52 (1.94, 3.54)	98.98 (0.38, 0.64)
	$r_P^*$	89.64 (5.10, 5.26)	94.76 (2.48, 2.76)	98.72 (0.70, 0.58)
(50,50)	$r_P$	89.96 (4.58, 5.46)	95.12 (2.26, 2.62)	99.0 (0.44, 0.56)
	$r_P^*$	89.72 (5.82, 4.46)	95.06 (2.74, 2.20)	98.7 (0.70, 0.60)

**Tabella 4.2:** Livelli di copertura empirica con  $\psi = 0.4$ ,  $\mu_X = 1.0489$ ,  $\sigma_X^2 = 1$ ,  $\mu_Y = 0$ ,  $\sigma_Y^2 = 1$  (std=0.4%).

La Tabella 4.3 mostra i risultati delle simulazioni ottenute considerando  $\psi = 0.6$ . Anche in questo caso, per campioni di dimensione inferiore a  $(m, n) = (30, 30)$ , la copertura degli intervalli di confidenza basati su  $r_P$  non raggiunge mai il valore nominale, mentre invece gli intervalli basati su  $r_P^*$  raggiungono i livelli di copertura desiderati.

Per quanto riguarda lo studio di simulazione nel caso  $\psi = 0.8$ , i cui risultati sono mostrati nella Tabella 4.4, l'intervallo basato su  $r_P$  presenta

(m,n)	Metodo	Livello di copertura nominale		
		90%	95%	99%
(5,5)	$r_P$	82.68 (2.31, 15.01)	90.03 (0.84, 9.12)	97.55 (0.02, 2.43)
	$r_P^*$	89.59 (4.22, 6.19)	95.50 (1.75, 2.75)	99.18 (0.28, 0.54)
(10,10)	$r_P$	85.30 (2.98, 11.72)	92.14 (1.64, 6.22)	97.76 (0.42, 1.82)
	$r_P^*$	89.54 (4.84, 5.62)	94.66 (2.34, 3.00)	98.52 (0.64, 0.84)
(20,20)	$r_P$	87.84 (3.34, 8.82)	93.74 (1.66, 4.60)	98.70 (0.38, 0.92)
	$r_P^*$	89.50 (4.82, 5.68)	94.62 (2.52, 2.86)	98.74 (0.64, 0.62)
(30,30)	$r_P$	88.92 (3.56, 7.52)	94.34 (1.72, 3.94)	98.94 (0.32, 0.74)
	$r_P^*$	89.90 (5.04, 5.06)	95.40 (2.40, 2.20)	98.84 (0.74, 0.42)
(50,50)	$r_P$	90.36 (3.72, 5.92)	95.28 (1.72, 3.00)	99.14 (0.28, 0.58)
	$r_P^*$	90.18 (5.06, 4.76)	95.20 (2.48, 2.32)	98.82 (0.60, 0.58)

**Tabella 4.3:** Livelli di copertura empirica con  $\psi = 0.6$ ,  $\mu_X = 1.6833$ ,  $\sigma_X^2 = 1$ ,  $\mu_Y = 0$ ,  $\sigma_Y^2 = 1$  (std=0.4%).

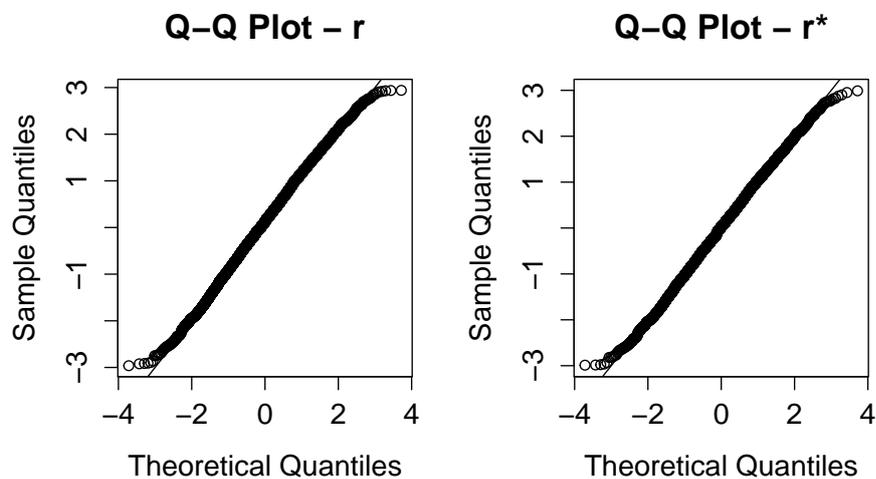
valori di copertura empirica inferiori a quelli nominali, in particolare con numerosità pari a  $(m, n) = (5, 5)$ ,  $(m, n) = (10, 10)$  e  $(m, n) = (20, 20)$ . L'intervallo basato su  $r_P^*$  per numerosità basse ha valori più vicini a quelli teorici.

In generale gli errori di copertura degli intervalli basati su  $r_P$  risultano un pò asimmetrici, soprattutto con basse numerosità, mentre quelli basati su  $r_P^*$  risultano abbastanza simmetrici.

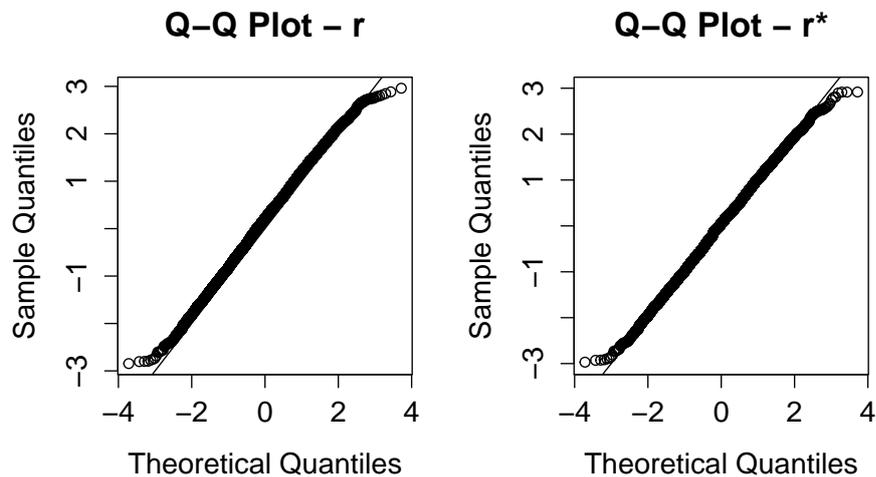
(m,n)	Metodo	Livello di copertura nominale		
		90%	95%	99%
(5,5)	$r_P$	84.54 (1.96, 13.5)	91.56 (1.02, 7.42)	98.22 (0.32, 1.46)
	$r_P^*$	90.86 (4.34, 4.8)	95.96 (1.92, 2.12)	99.20 (0.60, 0.20)
(10,10)	$r_P$	87.22 (2.74, 10.04)	93.26 (1.26, 5.48)	98.68 (0.12, 1.20)
	$r_P^*$	90.14 (4.70, 5.16)	94.78 (2.64, 2.58)	98.82 (0.62, 0.56)
(20,20)	$r_P$	88.32 (3.36, 8.32)	93.58 (1.60, 4.82)	98.40 (0.34, 1.26)
	$r_P^*$	89.52 (5.04, 5.44)	94.48 (2.76, 2.76)	98.54 (0.76, 0.70)
(30,30)	$r_P$	88.82 (3.26, 7.92)	94.26 (1.72, 4.02)	99.04 (0.28, 0.68)
	$r_P^*$	89.68 (5.16, 5.16)	94.56 (2.76, 2.68)	98.88 (0.64, 0.48)
(50,50)	$r_P$	89.66 (3.44, 6.9)	94.88 (1.58, 3.54)	98.90 (0.32, 0.78)
	$r_P^*$	89.98 (4.92, 5.1)	94.60 (2.52, 2.88)	98.62 (0.62, 0.76)

**Tabella 4.4:** Livelli di copertura empirica con  $\psi = 0.8$ ,  $\mu_X = 2.5632$ ,  $\sigma_X^2 = 1$ ,  $\mu_Y = 0$ ,  $\sigma_Y^2 = 1$  (std=0.4%).

Nelle Figure 4.1 e 4.2 vengono mostrati i QQ-Plot relativi ai valori di  $r_P$  e  $r_P^*$  ottenuti tramite simulazione nei casi in cui  $\psi = 0.4$  e  $\psi = 0.8$  con numerosità campionaria pari a  $(m, n) = (30, 30)$ . I grafici permettono di confrontare i quantili della distribuzione empirica con i quantili della distribuzione normale teorica. In entrambe i casi le osservazioni si dispongono lungo la diagonale, con lievi scostamenti nelle code, segno del fatto che l'approssimazione normale di  $r_P$  e  $r_P^*$  è adeguata.



**Figura 4.1:** QQ-plot dei valori di  $r_P$  ed  $r_P^*$  per  $\psi = 0.4$  e numerosità campionaria  $(m,n)=(30,30)$ .



**Figura 4.2:** QQ-plot dei valori di  $r_P$  ed  $r_P^*$  per  $\psi = 0.8$  e numerosità campionaria  $(m,n)=(30,30)$ .

### 4.3.2 Normali con varianze diverse

Il secondo studio di simulazione svolto è quello in cui  $X$  e  $Y$  hanno varianze diverse. Si analizza prima il caso in cui  $\sigma_X^2 = 3 > \sigma_Y^2 = 1$ , con  $\psi = (0.4, 0.6, 0.8)$ . I risultati delle simulazioni di questo caso sono riportati nelle Tabelle da 4.5 a 4.7.

Successivamente si tratterà il caso  $\sigma_X^2 = 0.5 < \sigma_Y^2 = 1$ , sempre con  $\psi = (0.4, 0.6, 0.8)$ .

(m,n)	Metodo	Livello di copertura nominale		
		90%	95%	99%
(5,5)	$r_P$	84.78 (2.96, 12.26)	90.97 (1.40, 7.63)	97.79 (0.08, 2.13)
	$r_P^*$	90.70 (3.19, 6.11)	95.94 (0.93, 3.13)	99.33 (0.04, 0.63)
(10,10)	$r_P$	87.38 (4.11, 8.51)	93.28 (2.09, 4.63)	98.8 (0.20, 1.00)
	$r_P^*$	90.63 (4.67, 4.70)	95.12 (2.57, 2.31)	99.1 (0.38, 0.52)
(20,20)	$r_P$	88.46 (4.56, 6.98)	93.40 (2.48, 4.12)	98.70 (0.42, 0.88)
	$r_P^*$	89.04 (5.74, 5.22)	94.08 (3.00, 2.92)	98.68 (0.66, 0.66)
(30,30)	$r_P$	88.80 (4.20, 7.00)	94.52 (1.98, 3.50)	98.98 (0.36, 0.66)
	$r_P^*$	89.64 (5.24, 5.12)	94.82 (2.60, 2.58)	98.82 (0.62, 0.56)
(50,50)	$r_P$	89.12 (4.70, 6.18)	94.8 (2.10, 3.10)	98.80 (0.42, 0.78)
	$r_P^*$	88.92 (5.78, 5.30)	94.4 (2.88, 2.72)	98.32 (0.82, 0.86)

**Tabella 4.5:** Livelli di copertura empirica con  $\psi = 0.4$ ,  $\mu_X = 1.2540$ ,  $\sigma_X^2 = 3$ ,  $\mu_Y = 0$ ,  $\sigma_Y^2 = 1$  (std=0.4%).

Le Tabelle 4.5 e 4.6 si riferiscono ai casi in cui  $\psi = 0.4$  e  $\psi = 0.6$ . Si può notare che gli intervalli basati su  $r_P$ , per numerosità pari e inferiori a  $(m, n) = (20, 20)$  forniscono un livello di copertura empirica lontano dai valori nominali. Gli intervalli basati su  $r_P^*$  si comportano invece meglio per campioni piccoli e sono quindi preferibili poiché più soddisfacenti in termini di copertura empirica.

(m,n)	Metodo	Livello di copertura nominale		
		90%	95%	99%
(5,5)	$r_P$	84.12 (3.18, 12.70)	91.01 (1.60, 7.39)	97.86 (0.221, 92)
	$r_P^*$	89.97 (4.73, 5.31)	94.99 (2.34, 2.66)	99.16 (0.34, 0.50)
(10,10)	$r_P$	87.26 (3.16, 9.58)	93.24 (1.64, 5.12)	98.34 (0.3, 1.36)
	$r_P^*$	90.54 (4.62, 4.84)	95.10 (2.36, 2.54)	98.94 (0.5, 0.56)
(20,20)	$r_P$	88.90 (3.48, 7.62)	93.60 (1.94, 4.46)	98.54 (0.4, 1.06)
	$r_P^*$	89.68 (4.88, 5.44)	94.36 (2.74, 2.90)	98.52 (0.7, 0.78)
(30,30)	$r_P$	88.66 (4.04, 7.30)	94.10 (1.84, 4.06)	98.88 (0.38, 0.74)
	$r_P^*$	88.94 (5.48, 5.58)	94.78 (2.60, 2.62)	98.60 (0.74, 0.66)
(50,50)	$r_P$	89.56 (3.94, 6.50)	94.8 (1.92, 3.28)	98.84 (0.42, 0.74)
	$r_P^*$	89.78 (4.96, 5.26)	94.5 (2.68, 2.82)	98.58 (0.68, 0.74)

**Tabella 4.6:** Livelli di copertura empirica con  $\psi = 0.6$ ,  $\mu_X = 2.1843$ ,  $\sigma_X^2 = 3$ ,  $\mu_Y = 0$ ,  $\sigma_Y^2 = 1$  (std=0.4%).

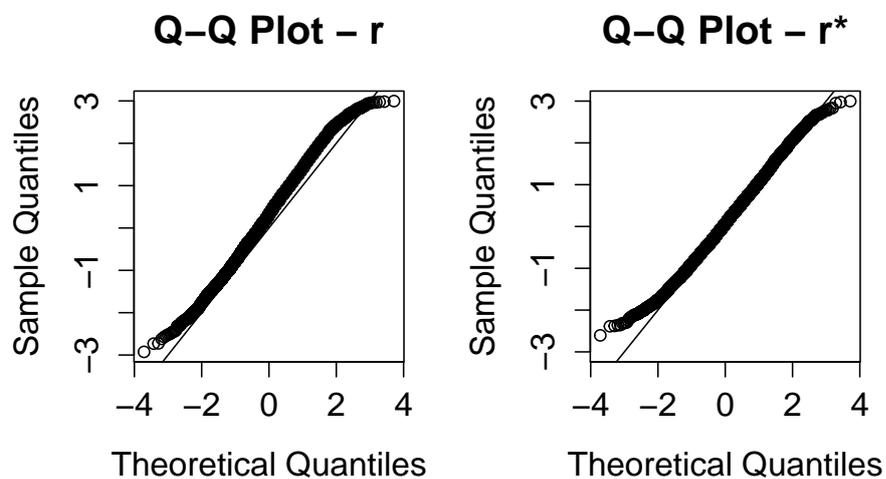
Dalla Tabella 4.7, che si riferisce al caso in cui  $\psi = 0.8$ , si può notare che

già ad  $(m, n) = (30, 30)$  gli intervalli basati su  $r_P$  non riescono a mantenere livelli di copertura empirica vicini a quelli teorici. Gli intervalli basati su  $r_P^*$  risultano quindi nuovamente preferibili rispetto ai primi.

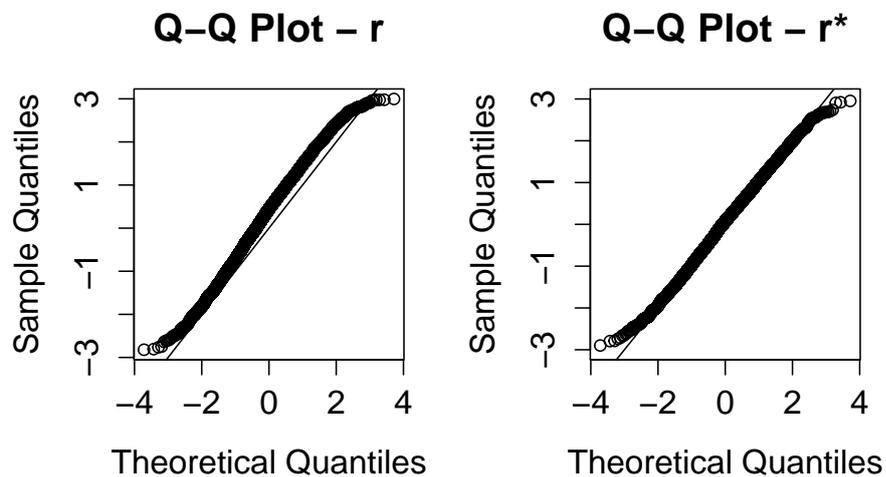
(m,n)	Metodo	Livello di copertura nominale		
		90%	95%	99%
(5,5)	$r_P$	83.10 (2.34, 14.56)	89.89 (1.38, 8.73)	97.58 (0.32, 2.10)
	$r_P^*$	89.45 (4.83, 5.73)	94.93 (2.42, 2.64)	99.02 (0.62, 0.36)
(10,10)	$r_P$	87.22 (2.30, 10.48)	92.94 (1.06, 6.00)	98.00 (0.14, 1.86)
	$r_P^*$	90.22 (4.26, 5.52)	94.90 (2.22, 2.88)	98.78 (0.62, 0.60)
(20,20)	$r_P$	88.44 (3.16, 8.40)	93.54 (1.64, 4.82)	98.72 (0.34, 0.94)
	$r_P^*$	89.42 (5.06, 5.52)	94.96 (2.50, 2.54)	98.92 (0.66, 0.42)
(30,30)	$r_P$	88.48 (3.60, 7.92)	94.3 (1.76, 3.94)	98.84 (0.26, 0.90)
	$r_P^*$	89.56 (5.36, 5.08)	94.7 (2.86, 2.44)	98.82 (0.64, 0.54)
(50,50)	$r_P$	89.66 (3.64, 6.70)	94.56 (1.7, 3.74)	98.80 (0.32, 0.88)
	$r_P^*$	90.06 (4.72, 5.22)	94.68 (2.5, 2.82)	98.64 (0.66, 0.70)

**Tabella 4.7:** Livelli di copertura empirica con  $\psi = 0.8$ ,  $\mu_X = 3.4247$ ,  $\sigma_X^2 = 3$ ,  $\mu_Y = 0$ ,  $\sigma_Y^2 = 1$  (std=0.4%).

Nelle Figure 4.3 e 4.4 vengono mostrati i QQ-Plot di  $r_P$  e  $r_P^*$  ottenuti tramite simulazione nei casi in cui  $\psi = 0.4$  e  $\psi = 0.6$  con numerosità campionaria pari a  $(m, n) = (5, 5)$ . Si nota nei grafici di sinistra delle Figure 4.3 e 4.4 un lieve scostamento dalla normalità per la distribuzione di  $r_P$ . Si specifica



**Figura 4.3:** QQ-plot dei valori di  $r_P$  ed  $r_P^*$  per  $\psi = 0.4$  e numerosità campionaria  $(m,n)=(5,5)$ .



**Figura 4.4:** QQ-plot dei valori di  $r_P$  ed  $r_P^*$  per  $\psi = 0.8$  e numerosità campionaria  $(m,n)=(5,5)$ .

a questo proposito che nel caso  $\psi = 0.4$ , per numerosità  $(m, n) = (5, 5)$ , si riscontra una certa instabilità numerica. Questa causa la mancata esecuzione della funzione `r.star` per alcuni campioni della simulazione (minore dell'1%, salvo il caso di varianze uguali con  $(m, n) = (5, 5)$ , dove raggiunge il 5%).

Rimanendo sempre nella situazione in cui  $X$  e  $Y$  hanno varianze diverse, si tratta ora il caso in cui  $\sigma_X^2 = 0.5 < \sigma_Y^2 = 1$ , con  $\psi = (0.4, 0.6, 0.8)$ . I risultati delle simulazioni sono riportati nelle Tabelle da 4.8 a 4.10.

(m,n)	Metodo	Livello di copertura nominale		
		90%	95%	99%
(5,5)	$r_P$	83.96 (2.73, 13.31)	90.96 (1.01, 8.03)	97.49 (0.08, 2.43)
	$r_P^*$	90.61 (3.22, 6.17)	96.03 (0.81, 3.16)	99.21 (0.08, 0.71)
(10,10)	$r_P$	87.71 (2.93, 9.37)	93.60 (1.26, 5.13)	98.40 (0.20, 1.40)
	$r_P^*$	90.71 (4.19, 5.09)	95.59 (1.93, 2.49)	98.96 (0.38, 0.66)
(20,20)	$r_P$	89.04 (3.88, 7.08)	94.30 (1.9, 3.80)	98.68 (0.48, 0.84)
	$r_P^*$	90.18 (4.94, 4.88)	95.22 (2.5, 2.28)	98.56 (0.68, 0.76)
(30,30)	$r_P$	88.84 (4.44, 6.72)	93.94 (2.30, 3.76)	98.90 (0.32, 0.78)
	$r_P^*$	89.40 (5.48, 5.12)	94.16 (2.94, 2.90)	98.64 (0.62, 0.74)
(50,50)	$r_P$	89.18 (4.48, 6.34)	94.56 (2.26, 3.18)	98.84 (0.56, 0.60)
	$r_P^*$	89.28 (5.44, 5.28)	94.50 (2.82, 2.68)	98.48 (0.72, 0.80)

**Tabella 4.8:** Livelli di copertura empirica con  $\psi = 0.4$ ,  $\mu_X = 0.8484$ ,  $\sigma_X^2 = 0.5$ ,  $\mu_Y = 0$ ,  $\sigma_Y^2 = 1$  (std=0.4%).

La Tabella 4.8 si riferisce al caso in cui  $\psi = 0.4$ . Si può notare che gli intervalli basati su  $r_P$  perdono in precisione a partire da campioni di numerosità  $(m, n) = (10, 10)$ , dove gli intervalli basati su  $r_P^*$  hanno invece coperture empiriche che raggiungono i livelli di copertura desiderati.

(m,n)	Metodo	Livello di copertura nominale		
		90%	95%	99%
(5,5)	$r_P$	83.25 (2.89, 13.87)	90.04 (1.60, 8.36)	97.41 (0.26, 2.32)
	$r_P^*$	90.14 (4.11, 5.75)	94.69 (2.32, 2.99)	99.20 (0.34, 0.46)
(10,10)	$r_P$	86.42 (3.20, 10.38)	92.48 (1.54, 5.98)	98.42 (0.18, 1.40)
	$r_P^*$	89.36 (4.96, 5.68)	94.82 (2.38, 2.80)	98.80 (0.52, 0.68)
(20,20)	$r_P$	88.42 (3.40, 8.18)	93.72 (1.82, 4.46)	98.72 (0.34, 0.94)
	$r_P^*$	89.72 (5.08, 5.20)	94.68 (2.64, 2.68)	98.66 (0.66, 0.68)
(30,30)	$r_P$	89.30 (3.36, 7.34)	94.18 (1.76, 4.06)	98.98 (0.20, 0.82)
	$r_P^*$	89.96 (4.74, 5.30)	94.64 (2.52, 2.84)	98.72 (0.58, 0.70)
(50,50)	$r_P$	89.34 (3.90, 6.76)	94.72 (1.84, 3.44)	98.98 (0.26, 0.76)
	$r_P^*$	89.74 (5.08, 5.18)	94.54 (2.82, 2.64)	98.72 (0.60, 0.68)

**Tabella 4.9:** Livelli di copertura empirica con  $\psi = 0.6$ ,  $\mu_X = 1.4071$ ,  $\sigma_X^2 = 0.5$ ,  $\mu_Y = 0$ ,  $\sigma_Y^2 = 1$  (std=0.4%).

La Tabella 4.9 mostra i risultati delle simulazioni ottenute considerando  $\psi = 0.6$ . Si nota che per numerosità pari e inferiori a  $(m, n) = (20, 20)$  i livelli di copertura degli intervalli di confidenza basati su  $r_P$  sono molto inferiori al

valore nominale fissato. Invece i livelli di copertura per gli intervalli basati su  $r_P^*$  sono sempre migliori e molto vicini ai livelli di copertura desiderati, anche per numerosità piccole.

(m,n)	Metodo	Livello di copertura nominale		
		90%	95%	99%
(5,5)	$r_P$	81.97 (2.32, 15.70)	89.75 (1.36, 8.89)	97.34 (0.38, 2.28)
	$r_P^*$	89.45 (5.01, 5.55)	94.77 (2.60, 2.62)	98.88 (0.78, 0.34)
(10,10)	$r_P$	86.60 (2.42, 10.98)	92.66 (1.42, 5.92)	98.24 (0.32, 1.44)
	$r_P^*$	90.36 (4.34, 5.30)	94.94 (2.42, 2.64)	98.74 (0.70, 0.56)
(20,20)	$r_P$	88.82 (2.78, 8.40)	94.40 (1.34, 4.26)	98.88 (0.20, 0.92)
	$r_P^*$	90.42 (4.70, 4.88)	95.04 (2.50, 2.46)	98.94 (0.56, 0.50)
(30,30)	$r_P$	89.1 (3.38, 7.52)	94.56 (1.6, 3.84)	98.86 (0.32, 0.82)
	$r_P^*$	89.9 (5.14, 4.96)	95.00 (2.6, 2.40)	98.74 (0.66, 0.60)
(50,50)	$r_P$	89.56 (3.56, 6.88)	94.86 (1.90, 3.24)	98.88 (0.36, 0.76)
	$r_P^*$	90.26 (4.94, 4.80)	94.70 (2.82, 2.48)	98.58 (0.68, 0.74)

**Tabella 4.10:** Livelli di copertura empirica con  $\psi = 0.8$ ,  $\mu_X = 2.1682$ ,  $\sigma_X^2 = 0.5$ ,  $\mu_Y = 0$ ,  $\sigma_Y^2 = 1$  (std=0.4%).

Per quanto riguarda lo studio di simulazione nel caso  $\psi = 0.8$ , i cui risultati sono mostrati nella Tabella 4.10, gli intervalli basati su  $r_P$  presentano valori di copertura empirica inferiori a quelli nominali, in particolare con

numerosità pari a  $(m, n) = (5, 5)$ ,  $(m, n) = (10, 10)$ . Gli intervalli basati su  $r_P^*$  per numerosità basse hanno invece valori più vicini a quelli teorici.

Nelle Figure 4.5 e 4.6 vengono mostrati i QQ-Plot relativi ai valori di  $r_P$  e  $r_P^*$  ottenuti tramite simulazione nei casi in cui  $\psi = 0.4$  e  $\psi = 0.6$  con numerosità campionaria pari a  $(m, n) = (10, 10)$ . I grafici permettono di confrontare i quantili della distribuzione empirica con i quantili della distribuzione normale teorica.

Nel primo caso le osservazioni si dispongono lungo la diagonale, con lievi scostamenti nelle code, segno del fatto che l'approssimazione normale di  $r_P$  e  $r_P^*$  è adeguata. Nel secondo grafico, l'approssimazione per  $r_P^*$  sembra essere migliore di quella per  $r_P$ .

## 4.4 Distribuzione gamma

Assumiamo ora che  $X$  e  $Y$ , che rappresentano rispettivamente gli ammalati e i sani, siano variabili casuali indipendenti e seguano le distribuzioni gamma rispettivamente del tipo  $Ga(\alpha_X, \beta_X)$  e  $Ga(\alpha_Y, \beta_Y)$ . Si considera il caso in cui  $\alpha_X\beta_X > \alpha_Y\beta_Y$ ; nel caso contrario è sufficiente invertire le notazioni  $X$  e  $Y$ . Si generano due campioni di dati  $x_1, \dots, x_m$  proveniente da  $X$  e  $y_1, \dots, y_n$  proveniente da  $Y$ .

In generale non esiste una formula esplicita per il *cut-off*  $c$  e deve essere quindi determinato numericamente. Nel caso particolare in cui  $\alpha = \alpha_X = \alpha_Y$  il *cut-point* si scrive (Schisterman, Perkins et al. 2005)

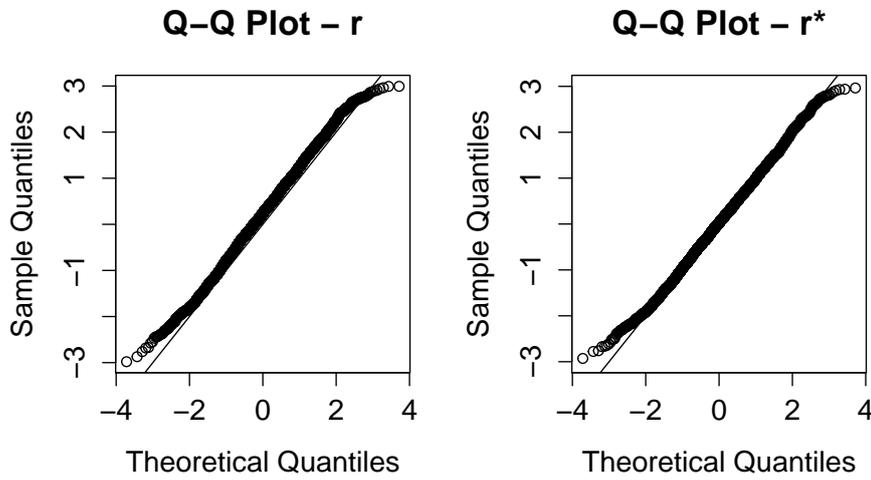
$$c = \alpha \times \ln \left( \frac{\beta_X}{\beta_Y} \right) \left( \frac{1}{\beta_Y} - \frac{1}{\beta_X} \right)^{-1}.$$

Nel caso particolare in cui  $\beta = \beta_X = \beta_Y$  il *cut-point* ha la forma

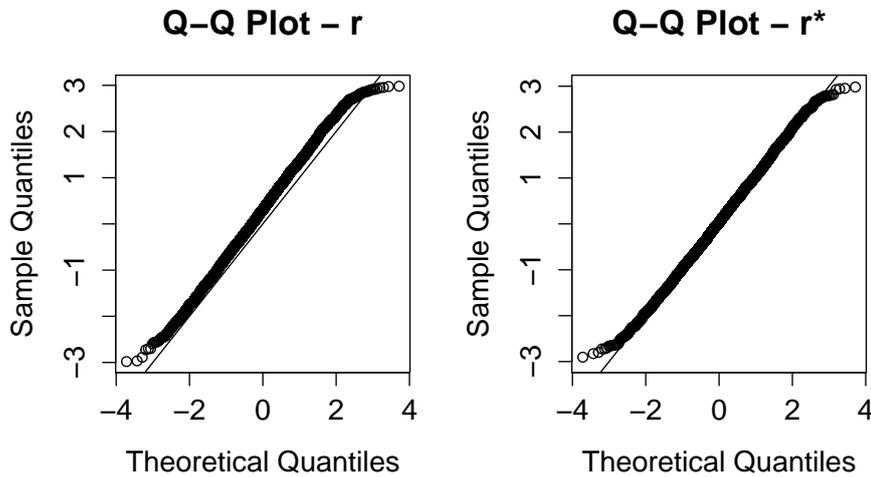
$$c = \beta \times \left( \frac{\Gamma(\alpha_X)}{\Gamma(\alpha_Y)} \right)^{\frac{1}{\alpha_X - \alpha_Y}}.$$

L'indice di Youden corrispondente si può esprimere come

$$\begin{aligned} \psi = \psi(\theta) &= \bar{F}(c) + G(c) - 1 \\ &= \int_c^{+\infty} \frac{e^{-x/\beta_X} x^{\alpha_X-1}}{\beta_X^{\alpha_X} \Gamma(\alpha_X)} dx + \int_0^c \frac{e^{-y/\beta_Y} y^{\alpha_Y-1}}{\beta_Y^{\alpha_Y} \Gamma(\alpha_Y)} dy - 1 \end{aligned}$$



**Figura 4.5:** QQ-plot dei valori di  $r_P$  ed  $r_P^*$  per  $\psi = 0.4$  e numerosità campionaria  $(m,n)=(10,10)$ .



**Figura 4.6:** QQ-plot dei valori di  $r_P$  ed  $r_P^*$  per  $\psi = 0.6$  e numerosità campionaria  $(m,n)=(10,10)$ .

con  $F$  e  $G$  funzioni di ripartizione di  $X$  e  $Y$ ,  $\bar{F} = 1 - F$  e  $c$  *cut-point* ottimale.

Sostituendo le stime di massima verosimiglianza  $\hat{\alpha}$ ,  $\hat{\beta}_X$ ,  $\hat{\beta}_Y$  nell'espressione precedente si ottengono le stime di  $c$  e  $\psi$ .

La riparametrizzazione nel caso in cui  $\alpha_X = \alpha_Y$  si scrive  $\theta = (\psi, \lambda)$ , dove  $\psi$  corrisponde all'indice di Youden e  $\lambda = (\alpha, \beta_Y)$ , con  $\alpha$  parametro di forma comune alle due popolazioni.

Nella seguente sezione si eseguiranno le simulazioni per studiare il caso gamma  $X$  e  $Y$  con parametri di forma uguali. Si assumerà che  $Y$  si distribuisca come una  $Ga \sim (1.5, 1)$ , quindi con  $\alpha = 1.5$  e  $\beta_Y = 1$ . I valori dei parametri di scala  $\beta_X$  della distribuzione di  $X$  verranno determinati in modo da ottenere un indice di Youden pari a  $J = (0.4, 0.6, 0.8)$ , come indicato nella Tabella 4.11.

Youden Index	Parametri di scala $\beta_X$
$J = 0.4$	0.4028
$J = 0.6$	0.2295
$J = 0.8$	0.1022

**Tabella 4.11:** Parametri di scala  $\beta_X$  con parametro di scala  $\alpha_X = 1.5$  e indici  $J$  di Youden fissati.

Le numerosità campionarie per  $X$  e  $Y$ , rispettivamente  $(m, n)$ , verranno fissate a  $(50, 50)$ ,  $(30, 30)$ ,  $(20, 20)$ ,  $(10, 10)$ ,  $(5, 5)$ . I livelli nominali di riferimento sono scelti al 90%, 95% e 99%.

Il numero di replicazioni è fissato a  $R = 5000$ . Per l'algoritmo Monte Carlo che calcola le covarianze empiriche il numero di ripetizioni è mantenuto al valore di default  $B = 1000$ .

I risultati delle simulazioni sono riportati nelle Tabelle da 4.12 a 4.14.

La Tabella 4.12 si riferisce al caso in cui  $\psi = 0.4$ . Si può notare che gli intervalli basati su  $r_P^*$  raggiungono i livelli di copertura desiderati per ogni livello di confidenza anche per campioni a bassa numerosità, dove invece quelli basati su  $r_P$  perdono in precisione. Le coperture empiriche mostrano infatti che gli intervalli basati su  $r$  sono buoni per numerosità pari o superiori a  $(m, n) = (20, 20)$ , mentre degradano per numerosità inferiori.

(m,n)	Metodo	Livello di copertura nominale		
		90%	95%	99%
(5,5)	$r_P$	88.16 (0.65, 11.19)	93.58 (0.00, 6.42)	97.95 (0.00, 2.05)
	$r_P^*$	94.13 (0.57, 5.30)	96.96 (0.08, 2.96)	99.12 (0.06, 0.82)
(10,10)	$r_P$	88.15 (3.97, 7.88)	93.93 (1.63, 4.43)	99.05 (0.0, 0.95)
	$r_P^*$	90.32 (4.54, 5.14)	95.75 (1.77, 2.48)	99.42 (0.1, 0.48)
(20,20)	$r_P$	88.18 (4.28, 7.54)	94.08 (1.90, 4.02)	98.94 (0.32, 0.74)
	$r_P^*$	89.54 (4.84, 5.62)	95.06 (2.14, 2.80)	98.94 (0.48, 0.58)
(30,30)	$r_P$	88.60 (4.88, 6.52)	94.24 (2.18, 3.58)	98.82 (0.32, 0.86)
	$r_P^*$	89.14 (5.34, 5.52)	94.22 (2.64, 3.14)	98.68 (0.52, 0.80)
(50,50)	$r_P$	89.72 (4.92, 5.36)	94.92 (2.40, 2.68)	99.14 (0.44, 0.42)
	$r_P^*$	89.22 (5.78, 5.00)	94.54 (3.02, 2.44)	98.60 (0.78, 0.62)

**Tabella 4.12:** Livelli di copertura empirica con  $\psi = 0.4$ ,  $\alpha_X = 1.5$ ,  $\beta_X = 1$ ,  $\alpha_Y = 1.5$ ,  $\beta_Y = 0.4028$  (std=0.4%).

La Tabella 4.13 mostra i risultati delle simulazioni ottenute considerando  $\psi = 0.6$ . Anche in questo caso, per campioni di dimensione inferiore a  $(m, n) = (30, 30)$ , la copertura degli intervalli di confidenza basati su  $r_P$  non raggiunge mai il valore nominale, mentre invece gli intervalli basati su  $r_P^*$  raggiungono i livelli di copertura desiderati.

Per quanto riguarda lo studio di simulazione nel caso  $\psi = 0.8$ , i cui risultati sono mostrati nella Tabella 4.14, l'intervallo basato su  $r_P$  presenta

(m,n)	Metodo	Livello di copertura nominale		
		90%	95%	99%
(5,5)	$r_P$	83.93 (3.86, 12.21)	90.68 (1.99, 7.33)	97.81 (0.16, 2.03)
	$r_P^*$	89.50 (5.06, 5.44)	94.90 (2.41, 2.69)	99.28 (0.20, 0.52)
(10,10)	$r_P$	87.12 (3.82, 9.06)	92.58 (2.04, 5.38)	97.88 (0.46, 1.66)
	$r_P^*$	89.06 (5.24, 5.70)	94.10 (2.92, 2.98)	98.48 (0.80, 0.72)
(20,20)	$r_P$	88.62 (3.96, 7.42)	94.00 (1.98, 4.02)	98.48 (0.48, 1.04)
	$r_P^*$	89.70 (4.92, 5.38)	94.44 (2.52, 3.04)	98.54 (0.72, 0.74)
(30,30)	$r_P$	89.50 (4.18, 6.32)	94.38 (2.16, 3.46)	98.80 (0.48, 0.72)
	$r_P^*$	89.92 (5.10, 4.98)	94.42 (2.84, 2.74)	98.86 (0.62, 0.52)
(50,50)	$r_P$	88.84 (4.9, 6.26)	94.28 (2.48, 3.24)	98.78 (0.46, 0.76)
	$r_P^*$	89.34 (5.6, 5.06)	94.32 (2.96, 2.72)	98.66 (0.66, 0.68)

**Tabella 4.13:** Livelli di copertura empirica con  $\psi = 0.6$ ,  $\alpha_X = 1.5$ ,  $\beta_X = 1$ ,  $\alpha_Y = 1.5$ ,  $\beta_Y = 0.2295$  (std=0.4%).

valori di copertura empirica inferiori a quelli nominali, in particolare con numerosità pari a  $(m, n) = (5, 5)$  e  $(m, n) = (10, 10)$ . L'intervallo basato su  $r_P^*$  per numerosità basse ha valori più vicini a quelli teorici.

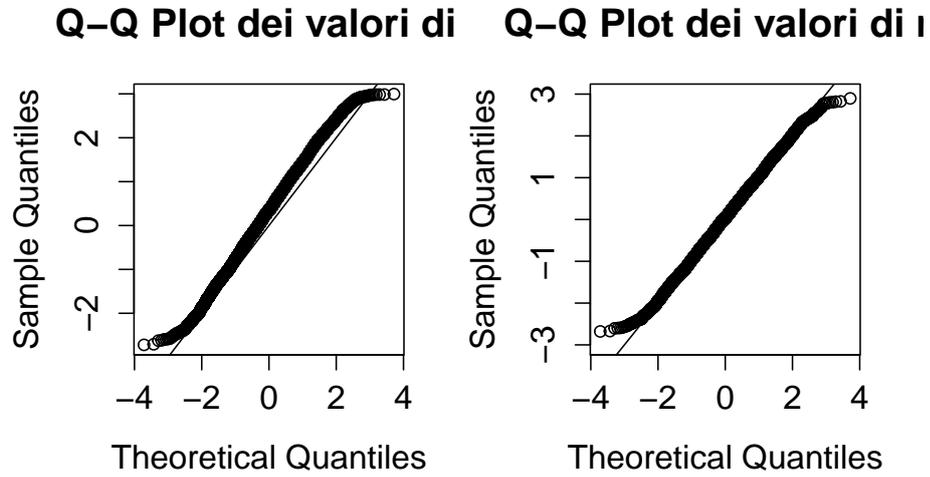
In generale gli errori di copertura degli intervalli basati su  $r_P$  risultano un pò asimmetrici, soprattutto con basse numerosità, mentre quelli basati su  $r_P^*$  risultano abbastanza simmetrici.

Nelle Figure 4.7 e 4.8 vengono mostrati i QQ-Plot relativi ai valori di  $r_P$  e

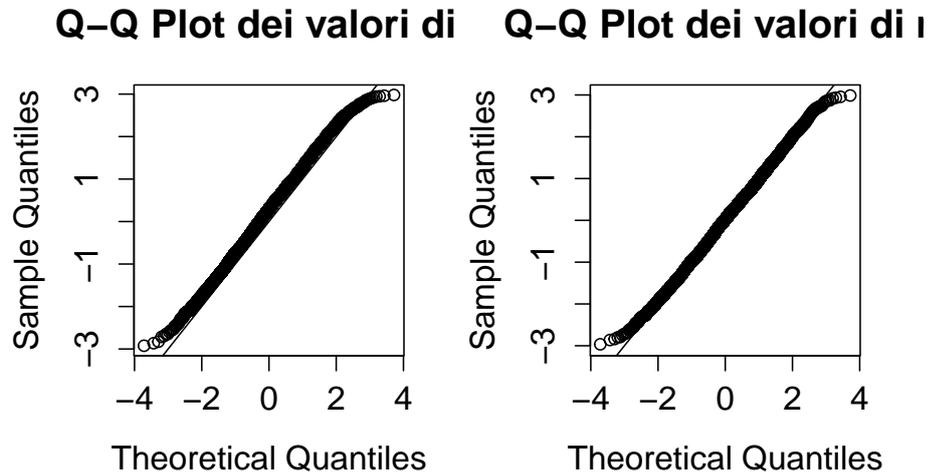
(m,n)	Metodo	Livello di copertura nominale		
		90%	95%	99%
(5,5)	$r_P$	82.80 (3.22, 13.98)	89.66 (1.64, 8.70)	96.94 (0.32, 2.74)
	$r_P^*$	88.98 (5.22, 5.80)	94.30 (2.76, 2.94)	98.64 (0.52, 0.84)
(10,10)	$r_P$	87.54 (2.98, 9.48)	93.48 (1.36, 5.16)	98.32 (0.26, 1.42)
	$r_P^*$	90.02 (4.86, 5.12)	95.02 (2.30, 2.68)	98.84 (0.48, 0.68)
(20,20)	$r_P$	88.26 (3.66, 8.08)	94.08 (1.64, 4.28)	98.74 (0.14, 1.12)
	$r_P^*$	89.66 (5.16, 5.18)	94.84 (2.60, 2.56)	98.76 (0.42, 0.82)
(30,30)	$r_P$	89.16 (4.30, 6.54)	94.48 (2.12, 3.40)	98.74 (0.50, 0.76)
	$r_P^*$	89.82 (5.44, 4.74)	94.48 (2.98, 2.54)	98.60 (0.76, 0.64)
(50,50)	$r_P$	89.2 (3.92, 6.88)	94.64 (1.80, 3.56)	99.16 (0.28, 0.56)
	$r_P^*$	89.1 (5.34, 5.56)	94.72 (2.64, 2.64)	98.72 (0.68, 0.60)

**Tabella 4.14:** Livelli di copertura empirica con  $\psi = 0.8$ ,  $\alpha_X = 1.5$ ,  $\beta_X = 1$ ,  $\alpha_Y = 1.5$ ,  $\beta_Y = 0.1022$  (std=0.4%).

$r_P^*$  ottenuti tramite simulazione nei casi in cui  $\psi = 0.6$  e  $\psi = 0.8$  con numerosità campionaria pari a  $(m, n) = (5, 5)$ . I grafici permettono di confrontare i quantili della distribuzione empirica con i quantili della distribuzione normale teorica. In entrambe i casi le osservazioni si dispongono lungo la diagonale, con lievi scostamenti nelle code, segno del fatto che l'approssimazione normale per la distribuzione di  $r_P$  è adeguata e quella di  $r_P^*$  è più accurata.



**Figura 4.7:** QQ-plot dei valori di  $r_P$  ed  $r_P^*$  per  $\psi = 0.6$  e numerosità campionaria  $(m,n)=(5,5)$ .



**Figura 4.8:** QQ-plot dei valori di  $r_P$  ed  $r_P^*$  per  $\psi = 0.8$  e numerosità campionaria  $(m,n)=(5,5)$ .

# Conclusione

L'obiettivo che si è perseguito in questa tesi è stato quello di approfondire lo studio delle approssimazioni di ordine superiore e capire in che misura la sua efficacia fosse vincolata alla numerosità campionaria. Si è quindi cercata una risposta attraverso uno studio di simulazione, confrontando il comportamento di due statistiche, la statistica radice con segno del log-rapporto di verosimiglianza  $r$ , che consente l'inferenza di verosimiglianza del primo ordine, e una sua versione modificata  $r^*$ , sviluppata nell'ambito della teoria asintotica di ordine superiore.

Si è fatto uso delle funzioni `rstar` e `rstar.ci` del pacchetto `LikelihoodAsy`, disponibile recentemente nella libreria di R. Esso permette infatti, tramite metodi di integrazione Monte Carlo, di implementare l'approssimazione di Skovgaard per il calcolo di  $r^*$ .

Il contesto dello studio è stato quello della stima dell'indice di Youden, sviluppato in ambito epidemiologico in relazione alla curva ROC. L'indice rappresenta una misura sintetica dell'efficacia di un test diagnostico. Ci si è focalizzati sulla valutazione della qualità degli intervalli di confidenza sull'indice, che costituisce il parametro di interesse.

Le simulazioni sono state effettuate per due schemi di campionamento, il primo con distribuzione normale e il secondo con distribuzione gamma. Si sono determinate le coperture empiriche degli intervalli di confidenza per i livelli nominali 90%, 95%, 99% e per le dimensioni campionarie  $(m, n) = (5, 5), (10, 10), (20, 20), (30, 30), (50, 50)$ . I valori dell'indice di Youden sono stati fissati a  $J = (0.4, 0.6, 0.8)$ . Si è inoltre valutata l'approssimazione gaussiana per la distribuzione delle statistiche  $r$  e  $r^*$ .

Dallo studio sono emersi risultati soddisfacenti: gli intervalli ottenuti con  $r^*$  si sono infatti dimostrati migliori di quelli ottenuti con  $r$  per piccole o

moderate numerosità campionarie.

Si è rilevato in particolare che

- per numerosità campionarie pari o minori di  $(m, n) = (20, 20)$  la copertura empirica degli intervalli di confidenza basati su  $r^*$  è più accurata della copertura degli intervalli di confidenza basati su  $r$ ;
- per numerosità campionarie pari o minori di  $(m, n) = (20, 20)$  la copertura empirica degli intervalli di confidenza basati su  $r^*$  raggiunge i livelli nominali, dove invece quella degli intervalli basati su  $r$  non raggiunge mai il livello di confidenza desiderato. Fanno eccezione le coperture nel caso di livello nominale 99%, che degradano solo per numerosità  $(m, n) = (5, 5)$ ;
- $r^*$  propone una migliore simmetria degli errori di non copertura bilaterali;
- per numerosità campionarie pari a  $(m, n) = (5, 5)$  e  $(m, n) = (10, 10)$  l'approssimazione alla normale di  $r_P^*$  sembra essere significativamente migliore di quella di  $r_P$ ;
- si è notata a volte un'instabilità numerica per  $J = 0.4$  nel caso di numerosità  $(m, n) = (5, 5)$ , che provoca la mancata esecuzione della funzione  $r^*$  nei cicli di simulazione (minore dell'1% dei casi, salvo per la distribuzione gamma con  $(m, n) = (5, 5)$  e normale di varianze uguali con  $(m, n) = (5, 5)$ , dove l'instabilità è più marcata).

Si può quindi concludere che i metodi di verosimiglianza di ordine superiore risultano preferibili su quelli del primo ordine nel caso di piccoli campioni e in presenza di parametri di disturbo, anche in modelli relativamente semplici come quelli considerati in questo elaborato.

# Bibliografia

- Azzalini, A. (2008). *Inferenza Statistica. Una presentazione basata sul concetto di verosimiglianza*. Milano: Springer Verlag.
- Bellio, R. e D.A. Pierce (2015a). *likelihoodAsy: Functions for Likelihood Asymptotics*. R package version 0.40. URL: <http://CRAN.R-project.org/package=likelihoodAsy>.
- Bellio, R. e D.A. Pierce (2015b). Modern Likelihood-Frequentist Inference. (Sottoposto per pubblicazione).
- Bellio, R. e D.A. Pierce (2015c). Practical likelihood asymptotics with the likelihoodAsy package. (Sottoposto per pubblicazione).
- Bottarelli, E. e S. Parodi (2003). Un approccio per la valutazione della validità dei test diagnostici: le curve ROC (Receiver Operating Characteristic). *Ann. Fac. Medic. Vet. di Parma* **23**, 49–68.
- Brazzale, A.R., A.C. Davison e N. Reid (2007). *Applied asymptotics: case studies in small-sample statistics*. Cambridge University Press.
- Pace, L. e A. Salvani (2001). *Introduzione alla statistica. II Inferenza, verosimiglianza, modelli*. Padova: Cedam.
- Schisterman, E.F. e N.J. Perkins (2007). Confidence intervals for the Youden index and corresponding optimal cut-point. *Communications in Statistics-Simulation and Computation* **36**, 549–563.
- Schisterman, E.F., N.J. Perkins, A. Liu e H. Bondell (2005). Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology* **16**, 73–81.
- Young, G.A. e R.L. Smith (2005). *Essentials of statistical inference*. Cambridge University Press.
- Zhou, H. (2008). New Non-Parametric Confidence Interval for the Youden. *Georgia State University*.



# Appendice A

## Codice R

### A.1 Caso normale con varianze uguali

---

#### Codice A.1: Funzioni per l'esecuzione di rstar e rstar.ci

```
### Funzione di log-verosimiglianza
llik.norm=function(theta,data) {
  mu1=theta[1]
  mu2=theta[2]
  omega=theta[3]
  lx=sum(dnorm(data$x,mean=mu1,sd=exp(omega),log=T))
  ly=sum(dnorm(data$y,mean=mu2,sd=exp(omega),log=T))
  return(lx+ly)
}

### Funzione generatrice di campioni
datagen.norm=function(theta,data) {
  nx=length(data$x)
  ny=length(data$y)
  data$x=rnorm(nx,mean=theta[1],sd=exp(theta[3]))
  data$y=rnorm(ny,mean=theta[2],sd=exp(theta[3]))
  data=list(x=data$x,y=data$y)
  return(data)
}
```

```

### Funzione di calcolo del parametro di interesse
tau.fun.psi=function(theta) {
  c=(theta[1]+theta[2])/2
  arg1=(theta[1]-c)/(exp(theta[3]))
  arg2=(c-theta[2])/(exp(theta[3]))
  psi=pnorm(arg1)+pnorm(arg2)-1
  if (psi < 1.0e-10) { psi = 1.0e-10 }
  if (psi > 1-1.0e-10) { psi = 1-1.0e-10 }
  tau=log(psi/(1-psi))
  return(tau)
}

### Valore di psi
psi=function(tau) exp(tau)/(1+exp(tau))

### Funzione rstar - stima di psi e calcolo di r e r*
rs<-rstar(data=dati,thetainit=c(1,1,1),floglik=llik.norm,
          fpsi=tau.fun.psi,psival=logit(psi),datagen=datagen.norm,
          R=1000,trace=FALSE,psidesc="Youden Index",constr.opt="alabama")

### Funzione rstar.ci - calcolo degli intervalli di confidenza di primo
ordine e di ordine superiore
rs.int<-rstar.ci(data=dati,thetainit=c(1,1,1),floglik=llik.norm,
                 fpsi=tau.fun.psi,datagen=datagen.norm,R=1000,
                 trace=FALSE,psidesc="Youden Index",constr.opt="alabama")

```

---

### Codice A.2: Studio di simulazione

---

```

# Numerosità campionaria
nx<-
ny<-
# Parametri delle distribuzioni normali
mu<-c(,)
sigma<-
# Dati
dati=list(x=rnorm(nx,mu[1],sigma),y=rnorm(ny,mu[2],sigma))

```

```
### Ciclo for per il calcolo di r e rstar (5000 replicazioni)
Nrep=100
r.out=NULL
rstar.out=NULL
for(i in 1:Nrep) {
  print(i)
  x=rnorm(nx,mu[1],sigma)
  y=rnorm(ny,mu[2],sigma)
  dati=list(x=x,y=y)
  res=try(rstar(data=dati,thetainit=c(1,1,1),floglik=llik.norm,
              fpsi=tau.fun.psi,psival=logit(psi),datagen=datagen.norm,
              R=1000,trace=FALSE,psidesc="Youden Index",
              constr.opt="alabama"))
  if(is.character(res)==T) next
  r.out[i]=res$r
  rstar.out[i]=res$rs
}
```

---

**Codice A.3:** Livelli percentuali di copertura empirica e errori bilaterali

---

```
# Livelli di copertura empirica per gli intervalli al 90%
cp1.r<-mean(abs(r.out)<qnorm(0.95))
cp1.rstar<-mean(abs(rstar.out)<qnorm(0.95))
e1.r<-c(mean(r.out<qnorm(0.05)),mean(r.out>qnorm(0.95)))
e1.rstar<-c(mean(rstar.out<qnorm(0.05)),mean(rstar.out>qnorm(0.95)))

# Livelli di copertura empirica per gli intervalli al 95%
cp2.r<-mean(abs(r.out)<qnorm(0.975))
cp2.rstar<-mean(abs(rstar.out)<qnorm(0.975))
e2.r<-c(mean(r.out<qnorm(0.025)),mean(r.out>qnorm(0.975)))
e2.rstar<-c(mean(rstar.out<qnorm(0.025)),mean(rstar.out>qnorm(0.975)))

# Livelli di copertura empirica per gli intervalli al 99%
cp3.r<-mean(abs(r.out)<qnorm(0.995))
cp3.rstar<-mean(abs(rstar.out)<qnorm(0.995))
e3.r<-c(mean(r.out<qnorm(0.005)),mean(r.out>qnorm(0.995)))
e3.rstar<-c(mean(rstar.out<qnorm(0.005)),mean(rstar.out>qnorm(0.995)))
```

```

### Matrici dei risultati
# Coperture empiriche
mat.cp<-matrix(c(cp1.r,cp2.r,cp3.r,cp1.rstar,cp2.rstar,cp3.rstar),2,3,
  byrow=T)
colnames(mat.cp)<-c("0.90","0.95","0.99")
rownames(mat.cp)<-c("cp.r","cp.r*")

# Errori bilaterali
mat.e<-matrix(c(e1.r,e2.r,e3.r,e1.rstar,e2.rstar,e3.rstar),2,6,byrow=T)
colnames(mat.e)<-c("0.90: low","upp","0.95: low","upp","0.99: low","upp")
rownames(mat.e)<-c("e.r","e.r*")

```

---

#### Codice A.4: qqplot di r e r\*

---

```

par(mfrow=c(1,2))
qqnorm(r.out[abs(r.out)<3], main="QQ Plot dei valori di r")
abline(0,1)
qqnorm(rstar.out[abs(rstar.out)<3], main="QQ Plot dei valori di r*",
  ylim=c(-3,3))
abline(0,1)

```

---

## A.2 Caso normale con varianze diverse

---

#### Codice A.5: Funzioni per l'esecuzione di rstar e rstar.ci

---

```

### Funzione di log-verosimiglianza
llik.norm=function(theta,data) {
  mu1=theta[1]
  omega1=theta[2]
  mu2=theta[3]
  omega2=theta[4]
  lx=sum(dnorm(data$x,mean=mu1,sd=exp(omega1),log=T))
  ly=sum(dnorm(data$y,mean=mu2,sd=exp(omega2),log=T))
  return(lx+ly)
}

```

```

### Funzione generatrice di campioni
datagen.norm=function(theta,data) {
  nx=length(data$x)
  ny=length(data$y)
  data$x=rnorm(nx,mean=theta[1],sd=exp(theta[2]))
  data$y=rnorm(ny,mean=theta[3],sd=exp(theta[4]))
  data=list(x=data$x,y=data$y)
  return(data)
}

### Funzione di calcolo del parametro di interesse
tau.fun.psi=function(theta) {
  a=theta[1]-theta[3]
  b=exp(theta[2])/exp(theta[4])
  c=(theta[3]*(b^2-1)-a+b*sqrt(a^2+(b^2-1)*exp(theta[4])*log(b^2)))/(b
    ^2-1)
  arg1=(theta[1]-c)/(exp(theta[2]))
  arg2=(c-theta[3])/exp(theta[4])
  psi=pnorm(arg1)+pnorm(arg2)-1
  if (psi < 1.0e-10) { psi = 1.0e-10 }
  if (psi > 1-1.0e-10) { psi = 1-1.0e-10 }
  tau=log(psi/(1-psi))
  return(tau)
}

### Valore di psi
psi=function(tau) exp(tau)/(1+exp(tau))

### Funzione di stima di psi e di calcolo di w, r e rstar
rs<-rstar(data=dati,thetainit=c(1,1,1,1),floglik=llik.norm,
  fpsi=tau.fun.psi,psival= logit(psi),datagen=datagen.norm,
  R=1000,trace=FALSE,psidesc="Youden Index",constr.opt="alabama")

### Funzione di calcolo degli intervalli di confidenza di primo ordine e
di ordine superiore
rs.int<-rstar.ci(data=dati,thetainit=c(1,1,1,1),floglik=llik.norm,

```

```
fpsi=tau.fun.psi,datagen=datagen.norm,R=1000,
trace=FALSE,psidesc="Youden Index",constr.opt="alabama")
```

---

### Codice A.6: Studio di simulazione

---

```
# Numerosità campionaria
nx<-
ny<-
# Parametri delle distribuzioni normali
mu<-c(,)
sigma<-c(,)
# Dati
dati=list(x=rnorm(nx,mu[1],sigma[1]),y=rnorm(ny,mu[2],sigma[2]))

### Ciclo for per il calcolo di r e rstar (5000 replicazioni)
Nrep=5000
r.out=NULL
rstar.out=NULL
for(i in 1:Nrep) {
  print(i)
  x=rnorm(nx,mu[1],sigma[1])
  y=rnorm(ny,mu[2],sigma[2])
  dati=list(x=x,y=y)
  res=try(rstar(data=dati,thetainit=c(1,1,1,1),floglik=llik.norm,
    fpsi=tau.fun.psi,psival= logit(psi),datagen=datagen.norm,
    R=1000,trace=FALSE,psidesc="Youden Index",
    constr.opt="alabama"))
  if(is.character(res)==T) next
  r.out[i]=res$r
  rstar.out[i]=res$rs
}
```

---

### Codice A.7: Livelli percentuali di copertura empirica e errori bilaterali

---

```
# Livelli di copertura empirica per gli intervalli al 90%
cp1.r<-mean(abs(r.out)<qnorm(0.95))
cp1.rstar<-mean(abs(rstar.out)<qnorm(0.95))
e1.r<-c(mean(r.out<qnorm(0.05)),mean(r.out>qnorm(0.95)))
```

```

e1.rstar<-c(mean(rstar.out<qnorm(0.05)),mean(rstar.out>qnorm(0.95)))

# Livelli di copertura empirica per gli intervalli al 95%
cp2.r<-mean(abs(r.out)<qnorm(0.975))
cp2.rstar<-mean(abs(rstar.out)<qnorm(0.975))
e2.r<-c(mean(r.out<qnorm(0.025)),mean(r.out>qnorm(0.975)))
e2.rstar<-c(mean(rstar.out<qnorm(0.025)),mean(rstar.out>qnorm(0.975)))

# Livelli di copertura empirica per gli intervalli al 99%
cp3.r<-mean(abs(r.out)<qnorm(0.995))
cp3.rstar<-mean(abs(rstar.out)<qnorm(0.995))
e3.r<-c(mean(r.out<qnorm(0.005)),mean(r.out>qnorm(0.995)))
e3.rstar<-c(mean(rstar.out<qnorm(0.005)),mean(rstar.out>qnorm(0.995)))

### Matrici dei risultati
# Coperture empiriche
mat.cp<-matrix(c(cp1.r, cp2.r, cp3.r, cp1.rstar, cp2.rstar, cp3.rstar), 2, 3,
               byrow=T)
colnames(mat.cp)<-c("0.90", "0.95", "0.99")
rownames(mat.cp)<-c("cp.r", "cp.r*")

# Errori bilaterali
mat.e<-matrix(c(e1.r, e2.r, e3.r, e1.rstar, e2.rstar, e3.rstar), 2, 6, byrow=T)
colnames(mat.e)<-c("0.90: low", "upp", "0.95: low", "upp", "0.99: low", "upp")
rownames(mat.e)<-c("e.r", "e.r*")

```

---

### Codice A.8: qqplot di r e r\*

---

```

par(mfrow=c(1,2))
qqnorm(r.out[abs(r.out)<3], main="Q-Q Plot dei valori di r")
abline(0,1)
qqnorm(rstar.out[abs(rstar.out)<3], main="Q-Q Plot dei valori di r*",
        ylim=c(-3,3))
abline(0,1)

```

---

## A.3 Caso gamma

### Codice A.9: Funzioni per l'esecuzione di rstar e rstar.ci

```
### Funzione di log-verosimiglianza
llik.gam=function(theta,data) {
  alpha=theta[1]
  beta1=theta[2]
  beta2=theta[3]
  lx=sum(dgamma(data$x,shape=exp(alpha),scale=exp(beta1),log=T))
  ly=sum(dgamma(data$y,shape=exp(alpha),scale=exp(beta2),log=T))
  return(lx+ly)
}

### Funzione generatrice di campioni
datagen.gam=function(theta,data) {
  nx=length(data$x)
  ny=length(data$y)
  data$x=rgamma(nx,shape=exp(theta[1]),scale=exp(theta[2]))
  data$y=rgamma(ny,shape=exp(theta[1]),scale=exp(theta[3]))
  data=list(x=data$x,y=data$y)
  return(data)
}

### Funzione di calcolo del parametro di interesse
tau.fun.psi=function(theta) {
  cN=exp(theta[1])*(theta[3]-theta[2])
  cD=1/exp(theta[2])-1/exp(theta[3])
  c=cN/cD
  arg1=pgamma(c,shape=exp(theta[1]),scale=exp(theta[3]))
  arg2=pgamma(c,shape=exp(theta[1]),scale=exp(theta[2]))
  psi=1-arg1+arg2-1
  if (psi < 1.0e-10) {psi = 1.0e-10}
  if (psi > 1-1.0e-10) {psi = 1-1.0e-10}
  tau=log(psi/(1-psi))
  return(tau)
}
```

```

### Valore di psi
psi=function(tau) exp(tau)/(1+exp(tau))

### Funzione di stima di psi e di calcolo di w, r e rstar
rs<-rstar(data=dati, thetainit=c(1,1,1), floglik=llik.gam,
          fpsi=tau.fun.psi, psival= -0.405, datagen=datagen.gam,
          R=1000, trace=FALSE,psidesc="Youden Index",constr.opt="
          alabama")

### Funzione di calcolo degli intervalli di confidenza di primo ordine e
di ordine superiore
rs.int<-rstar.ci(data=dati, thetainit=c(1,1,1), floglik=llik.gam,
                fpsi=tau.fun.psi, datagen=datagen.gam,
                R=1000, trace=FALSE,psidesc="Youden Index",
                constr.opt="alabama")

```

---

### Codice A.10: Studio di simulazione

---

```

# Numerosità campionaria
nx<-
ny<-
# Parametri delle distribuzioni normali
alpha<-
beta<-c(,)
# Dati
dati=list(x=rgamma(nx,alpha,beta[1]),y=rgamma(ny,alpha,beta[2]))

### Ciclo for per il calcolo di r e rstar (5000 repliche)
Nrep=5000
r.out=NULL
rstar.out=NULL
for(i in 1:Nrep) {
  print(i)
  x=rgamma(nx,shape=alpha,scale=beta[1])
  y=rgamma(ny,shape=alpha,scale=beta[2])
  dati=list(x=x,y=y)
}

```

```

res=try(rstar(data=dati,thetainit=c(1,1,1),floglik=llik.gam,
            fpsi=tau.fun.psi,psival=logit(psi),datagen=datagen.gam,
            R=1000,trace=FALSE,psidesc="Youden Index",
            constr.opt="alabama"))
if(is.character(res)==T) next
r.out[i] = res$r

rstar.out[i] = res$rs
}

```

### Codice A.11: Livelli percentuali di copertura empirica e errori bilaterali

```

# Livelli di copertura empirica per gli intervalli al 90%
cp1.r<-mean(abs(r.out)<qnorm(0.95))
cp1.rstar<-mean(abs(rstar.out)<qnorm(0.95))
e1.r<-c(mean(r.out<qnorm(0.05)),mean(r.out>qnorm(0.95)))
e1.rstar<-c(mean(rstar.out<qnorm(0.05)),mean(rstar.out>qnorm(0.95)))

# Livelli di copertura empirica per gli intervalli al 95%
cp2.r<-mean(abs(r.out)<qnorm(0.975))
cp2.rstar<-mean(abs(rstar.out)<qnorm(0.975))
e2.r<-c(mean(r.out<qnorm(0.025)),mean(r.out>qnorm(0.975)))
e2.rstar<-c(mean(rstar.out<qnorm(0.025)),mean(rstar.out>qnorm(0.975)))

# Livelli di copertura empirica per gli intervalli al 99%
cp3.r<-mean(abs(r.out)<qnorm(0.995))
cp3.rstar<-mean(abs(rstar.out)<qnorm(0.995))
e3.r<-c(mean(r.out<qnorm(0.005)),mean(r.out>qnorm(0.995)))
e3.rstar<-c(mean(rstar.out<qnorm(0.005)),mean(rstar.out>qnorm(0.995)))

### Matrici dei risultati
# Coperture empiriche
mat.cp<-matrix(c(cp1.r,cp2.r,cp3.r,cp1.rstar,cp2.rstar,cp3.rstar),2,3,
              byrow=T)
colnames(mat.cp)<-c("0.90","0.95","0.99")
rownames(mat.cp)<-c("cp.r","cp.r*")

```

```
# Errori bilaterali
mat.e<-matrix(c(e1.r,e2.r,e3.r,e1.rstar,e2.rstar,e3.rstar),2,6,byrow=T)
colnames(mat.e)<-c("0.90: low","upp","0.95: low","upp","0.99: low","upp")
rownames(mat.e)<-c("e.r","e.r*")
```

---

**Codice A.12:** qqplot di r e r\*

---

```
par(mfrow=c(1,2))
qqnorm(r.out[abs(r.out)<3], main="QQ Plot dei valori di r")
abline(0,1)
qqnorm(rstar.out[abs(rstar.out)<3], main="QQ Plot dei valori di r*",
       ylim=c(-3,3))
abline(0,1)
```

---

