Indice

Introduzione

Lo scopo di questa tesi è duplice. Da un lato, prettamente applicativo, si desidera presentare un’analisi statistica di un insieme di dati reali, avente come obiettivo la validazione di un nuovo test diagnostico. In sostanza, si vogliono presentare analisi e test statistici che permettono di valutare la bontà di una tecnica sperimentale. Da un punto di vista metodologico, si desidera presentare e illustrare un’applicazione di tecniche di inferenza più avanzate. Queste sono state recentemente illustrate con casi di studio nel testo di Brazzale *et al.* (2007) e consistono in uno sviluppo di metodi asintotici per quantità di verosimiglianza, utili soprattutto in casi in cui la numerosità campionaria è ridotta.
I dati utilizzati in questo studio provengono dal database dell’Azienda Ospedaliera di Padova e sono stati selezionati da Silvia Galante, laureanda in Medicina e Chirurgia dell’Università degli Studi di Padova, in collaborazione con il Professore Ambrogio Fassina.

Il dataset esaminato è composto da 28 pazienti. Questi pazienti sono soggetti ricoverati in momenti, reparti e per motivazioni diverse presso l’Azienda Ospedaliera di Padova. La caratteristica comune, grazie alla quale è avvenuta la loro selezione, è l’osservazione di lesioni granulomatose nei linfonodi durante l'esame microscopico. Per tali pazienti si sospetta, quindi, la presenza di un micobatterio che ne avrebbe causato la successiva infezione. Obiettivo di questo studio è valutare le analisi diagnostiche eseguite per individuare la patologia in oggetto, sperimentando nuovi test diagnostici al fine di valutarne l'efficacia e le tempistiche. Sui pazienti sono state eseguite delle analisi sui linfonodi conservati in paraffina, allo scopo di approfondire la conoscenza dello stato del linfonodo e per rafforzare l'ipotesi iniziale. Le analisi classiche che sono state eseguite sono due. La prima è l'**esame microscopico diretto**, ovvero l'osservazione della composizione e delle caratteristiche del linfonodo in via esplorativa e la successiva colorazione Ziehl-Neelsen, che presenta una colorazione rossa in presenza di micobatteri. La seconda è la **coltura**, che attualmente costituisce il test più affidabile che, oltre ad individuare la presenza di micobatteri, ne determina anche la tipologia. L'unico svantaggio della cultura è che i tempi di refertazione sono da 7-15 giorni a 3-6 settimane. In questo contesto, lo scopo di questo lavoro è valutare l'efficacia di una nuova tecnica diagnostica, chiamata **PCR** (*Polimerase Chain Reaction*), per verificare la presenza di micobatteri nei linfonodi. Tale tecnica permetterebbe di recuperare risposte attendibili in tempi più brevi rispetto agli esami classici.

Il Capitolo 1 è dedicato alla presentazione della patologia. Vengono presentate le caratteristiche principali delle infezioni da micobatteri: le cause principali, i sintomi più comuni e le terapie prescritte in caso di accertata presenza dell'infezione. Viene presa come esempio la tubercolosi e viene presentata la patologia descrivendone, in aggiunta, i risvolti fisici sui pazienti.
Di seguito vengono presentati i due test normalmente applicati in fase diagnostica, ossia l'esame microscopico diretto nel linfonodo e l'esame più approfondito, la cultura. Il primo rappresenta una prima esplorazione del linfonodo, per valutarne le caratteristiche e la consistenza; il secondo, più accurato, dichiara in maniera affidabile e decisamente più sicura se il paziente ha un'infezione da *Mycobacterium*, specificandone anche la tipologia. Nel capitolo vengono anche richiamati i concetti di **sensibilità** e **specificità** di un test diagnostico, che ne

permettono di valutare l'affidabilità. Per la sensibilità e la specificità vengono anche fornite le espressioni per il calcolo di intervalli di confidenza, facendo riferimento anche a risultati più recenti e accurati rispetto al classico intervallo alla Wald per una proporzione.

Il Capitolo 2 entra più in concreto nel caso di studio di Padova, descrivendo la struttura del dataset, anticipata brevemente nel Capitolo 1, analizzando le variabili e mettendone in risalto le particolarità. Vengono anche studiate le relazioni tra le diverse variabili, cercando di individuare legami significativi e caratteristiche comuni.

Il Capitolo 3 si differenzia dai precedenti, essendo un capitolo di rassegna su risultati recenti relativi alla teoria della verosimiglianza e sui metodi asintotici (HOA, *higher-order asymptotics*). Si parte descrivendo le tecniche inferenziali classiche di base per l'inferenza, sottolineando i risultati di distribuzione del primo ordine, fino ad arrivare alle moderne procedure metodologiche, dette di ordine superiore. Queste tecniche sono illustrate nel caso particolare delle famiglie esponenziali.

Nel Capitolo 4 viene, infine, descritta l'applicazione delle procedure di inferenza di primo ordine e di ordine superiore per il caso di studio. In principio vengono calcolati gli indici di sensibilità e specificità, specificando sia le stime puntuali che intervallari per le variabili più importanti, ossia l'analisi microscopica, la PCR ed infine la coltura. Inoltre, vengono discusse delle analisi sul log-rapporto delle quote in tabelle 2x2, create dal confronto tra le variabili.

In sintesi, gli obiettivi di questa tesi sono due:

1. studiare le nuove tecniche diagnostiche e valutare se possono essere utilizzate in sostituzione (o congiuntamente) agli usuali metodi;

2. sfruttare recenti procedure di inferenza per calcolare intervalli di confidenza per sensibilità e specificità degli esami, che forniscono risultati più attendibili con numerosità campionarie piccole; presentare metodi asintotici di ordine elevato nell'analisi di tabelle di contingenza 2x2.

Per le analisi statistiche presentate nella tesi è stato utilizzato il programma statistico R. Una copia di R può essere scaricata gratuitamente accedendo all'indirizzo Web:

http://www.r-project.org/,

in cui si trovano versioni del linguaggio per diversi sistemi operativi (MS-Windows, Unix, Linux). In questa tesi si è utilizzata la versione R.2.7.0 per Windows. Per la visualizzazione di alcuni comandi utilizzati si rimanda all'Appendice, in cui sono riportati nel dettaglio tutti i codici.

Capitolo 1

Il caso di studio:
Le infezioni da *Mycobacterium*

Obiettivo di questo capitolo è presentare la patologia, e il corrispondente dataset, oggetto di studio in questa tesi, la cui identificazione avviene tramite la valutazione della presenza di micobatteri nei linfonodi analizzati. Nello specifico, vengono anche illustrate le tecniche diagnostiche utilizzate per l'identificazione della malattia.

Il capitolo si apre con uno sguardo generale sulle infezioni da *Mycobacterium* (MB) (vedi Besana *et al.*, 1995; Moroni *et al.*, 2002 e Murray *et al.* 2003). In particolare, viene preso in esame il micobatterio della Tubercolosi (MTB) per sottolineare le caratteristiche principali di tali infezioni e individuarne i sintomi il prima possibile, allo scopo di assegnare prontamente la terapia adeguata (cfr. sito internet Giunta Regionale, 2007).

Successivamente, si procede alla spiegazione delle fasi con cui il personale medico formula la diagnosi di un paziente. Sono richiamate, inoltre, le definizioni di sensibilità e specificità di un test diagnostico.

Infine, si descrivono le analisi classiche utilizzate per individuare il principio di infezione dovuta a micobatteri, ovvero l'esame microscopico diretto e la coltura. Sarà anche introdotta una nuova tecnica in fase di sperimentazione, ossia la ***Polimerase Chain Reaction*** (o PCR).

I dati sono stati raccolti dal Prof. Ambrogio Fassina, Direttore del Laboratorio di Citodiagnostica dell'Azienda Ospedaliera di Padova e Docente di Anatomia Patologica nel corso di Laurea in Medicina e Chirurgia dell'Università degli Studi di Padova. Successivamente i dati sono stati integrati dalla laureanda Silvia Galante, con il sostegno dei responsabili dei vari laboratori, con i risultati degli esami svolti sui linfonodi dei pazienti conservati in paraffina.

1.1 Le infezioni da *Mycobacterium*

A distanza di oltre 120 anni dalla scoperta del ***Mycobacterium Tuberculosis*** (MTB), e nonostante i progressi compiuti in ambito diagnostico e terapeutico, la **tubercolosi** (TB) rappresenta ancora nel mondo la principale causa di morte da singolo agente infettivo.

Le linee guida per il controllo della tubercolosi pubblicato dalla Giunta Regionale nel 2007 riportano uno studio in cui circa un terzo della popolazione mondiale, cioè due miliardi di persone, sia stato infettato dal micobatterio della tubercolosi. Da otto a dieci milioni di persone sviluppano ogni anno una tubercolosi attiva e circa un quarto di queste muore a causa della malattia. Oltre il 90% dei casi e dei decessi si verifica nei paesi in via di sviluppo. Inoltre, una minaccia preoccupante è costituita dalla crescente resistenza a vari farmaci.

La tubercolosi ha tre principali caratteristiche:

1) è contagiosa;
2) è cronica;
3) è causata dal MTB.

La tubercolosi colpisce prevalentemente i polmoni, ma può colpire anche altri organi e si manifesta con una reazione infiammatoria granulomatosa che, però, non è esclusiva della tubercolosi. Infatti, tale reazione può derivare anche da altri tipi di infezioni.

I pilastri di un programma di controllo della tubercolosi sono costituiti da:

a) diagnosi precoce,
b) trattamento adeguato e tempestivo,
c) corretto follow-up,
d) prevenzione,
e) sorveglianza delle resistenze,
f) sorveglianza epidemiologica.

Un'efficace strategia di controllo della tubercolosi richiede la disponibilità di un'efficiente rete di laboratori di diagnostica dell'infezione tubercolare e delle micobatteriosi in generale.

Agenti causali dell'infezione tubercolare sono alcune specie di micobatteri raggruppati sotto la denominazione di *Mycobacterium Tuberculosis Complex* (MTC).

Una carattestica di tutti i micobatteri è che sono composti da bacilli sottili, *aerobi*[1] e *asporigeni*[2]. Inoltre, essi sono resistenti all'acido-alcol, diventano rossi quando sono sottoposti alla colorazione di Ziehl-Neelsen e non sono sensibili, invece, alla colorazione di Gram. Infatti, la natura cerosa dell'involucro esterno del micobatterio lo rende altamente impermeabile ai coloranti ordinari e si deve, perciò, ricorrere alla colorazione di Ziehl-Neelsen. Quest'ultima viene effettuata con versamento di Fucsia basica sul vetrino (tale reagente ha carica positiva che gli conferisce affinità per strutture acide quali la superficie dei micobatteri), si lascia evaporare scaldandolo con fiamma per poi essere lavato e decolorato con

[1] *Aerobi*: organismi che utilizzano l'ossigeno dell'aria e dell'acqua per produrre energia necessaria alle funzioni vitali.
[2] *Asporigeni:* non avere la fase di spora; che è una fase del ciclo vitale di alcuni batteri.

alcool-acido fino alla scomparsa del colorante. Poi, si rilava con l'acqua e si contrasta il risultato con il Blu di metilene. Infine, si risciacqua nuovamente. Gli organismi acido-resistenti (quali i micobatteri), appaiono colorati di rosso, mentre i non acido-resistenti di blu.

Ci sono due tipi di trasmissione della tubercolosi: **congenita** (ad esempio una trasmissione da madre a feto) o **acquisita**. Il MTB si trasmette quasi esclusivamente con la seconda tipologia di trasmissione, per contagio interumano, che può avvenire:

1. per via aerea, ossia attraverso goccioline di saliva, soprattutto con la tosse, dall'individuo affetto da tubercolosi bacillifera polmonare, bronchiale, tracheale o laringea;
2. per via gastro-intestinale, ad esempio per ingestione di latte contaminato da *M.bovis*;
3. per via ematolinfatica, ossia attraverso batteri penetrati per via gastro-intestinale e che raggiungono il polmone o altri organi.

La trasmissione viene facilitata negli ambienti affollati e poco aerati.

Poiché in media solo il 30%-40% dei contatti stretti di un caso di TB bacillifera viene infettato, si ritiene che un'immunità congenita protegga certi soggetti dall'infezione.

Nel soggetto infettato si possono presentare due situazioni di malattia:

A) **malattia tubercolare**, generata dal prevalere dei fattori aggressivi, cioè la carica microbica e la sua virulenza, su quelli difensivi, rappresentati dal sistema immunitario dell'ospite. Dati epidemiologici indicano che circa il 10% dei soggetti infettati sviluppa una tubercolosi, metà entro due anni dall'infezione e metà in un momento successivo della vita.

B) **infezione tubercolare latente** (ITBL), condizione che risulta dalla capacità del sistema immunitario dell'ospite di opporsi all'evolversi dell'infezione. Questa condizione può durare per tutta la vita, ma l'equilibrio può rompersi per il verificarsi di stati di deficienza immunitaria, anche transitoria.

Dal momento dell'infezione (intesa come penetrazione del bacillo nell'organismo), al momento dello sviluppo di una reazione positiva alla tubercolina, può trascorrere un tempo variabile dalle 2 alle 12 settimane. Il rischio di malattia è più elevato nei 6 mesi dopo l'infezione e resta elevato per circa due anni.

I sintomi di sospetto della malattia sono:

- febbre, soprattutto *serotina*[3];
- sudorazione notturna;
- calo ponderale;
- *astenia*[4];
- inappetenza;
- tosse produttiva e persistente (il sintomo più comune di TB polmonare).

Altri possibili sintomi di allarme di TB polmonare sono:

- il dolore toracico, spesso dovuto a concomitante (cioè che si manifesta insieme con altri fenomeni),
- interessamento *pleurico*[5],
- l'*emoftoe*[6].

Questi sintomi non sono specifici, ma, come per i sintomi elencati in precedenza, la possibilità di una TB va sempre tenuta presente.

[3] *Serotina*: tardiva, che matura più tardi.
[4] *Astenia*: riduzione della forza muscolare, per cui i movimenti sono eseguiti con scarsa energia, anche se sono tutti possibili e completi.
[5] *Pleurico*: della membrana sierosa che riveste il polmone.
[6] *Emoftoe*: emissione con la tosse di sangue proveniente dalle vie aeree misto a catarro. Può essere causata da banali infiammazioni respiratorie, raramente può rappresentare sintomo di carcinoma polmonare.

1.2 Come elaborare un test diagnostico

Quando un paziente si presenta in ospedale per essere curato, il principale obiettivo del medico è formulare la diagnosi, per procedere poi con la definizione della cura alla malattia.

La formulazione di una diagnosi è un processo complesso in quanto, oltre a valutare i test diagnostici, sintomi, segni e risultati degli esami di laboratorio, si basa anche sul giudizio soggettivo: il cosiddetto occhio clinico del medico.

Obiettivo di questa tesi è testare un nuovo test diagnostico per la TB, la PCR, proprio in questa fase di valutazione della patologia del paziente. In generale, per test diagnostico si intende una qualunque procedura utile all'identificazione di uno stato di malattia. Gli esiti di un test diagnostico dicotomico possono essere:

- positivo, che induce a sospettare la presenza della malattia;
- negativo, che sembra escluderne la presenza.

L'affidabilità di un test diagnostico è generalmente valutata in termini di **sensibilità** e **specificità**. Per presentare la sensibilità e la specificità è utile partire da una tabella a doppia entrata (vedi Tabella 1.1) che classifica gli *n* pazienti nello studio in positivi e negativi al test diagnostico rispetto alla presenza o assenza della malattia. In particolare, nella Tabella 1.1 i valori concordanti sono rappresentati da: **VN** che denota i veri negativi (ovvero l'insieme dei soggetti che non hanno la malattia e che hanno avuto esito negativo al test) e **VP** che indica i veri positivi (cioè i pazienti che hanno la malattia e il test ha avuto risultato positivo). I valori discordanti sono, invece, indicati da: **FN** falsi negativi (hanno la malattia ma il risultato del test era negativo) e **FP** falsi positivi (i soggetti sono sani anche se il test è risultato positivo).

		Malattia		
Esito test		**Assente**	**Presente**	**Totale**
	Negativo	**VN**	**FN**	**NEGATIVI**
	Positivo	**FP**	**VP**	**POSITIVI**
	Totale	**SANI** (n_{SP})	**MALATI** (n_{SN})	$\boldsymbol{n}$

Tabella 1.1 *Tabella di contingenza 2x2.*

Un buon test diagnostico tende a fornire esiti positivi in soggetti che presentano la malattia. La probabilità che un test ha di fornire esiti positivi nei malati, prende il nome di **sensibilità** (SN). La probabilità che un test diagnostico ha di fornire esiti negativi nei pazienti non malati prende, invece, il nome di **specificità** (SP).
Queste due quantità vengono calcolate a partire dai dati osservati, rispettivamente, come:

$$SN = \frac{VP}{FN + VP} \text{ e } SP = \frac{VN}{VN + FP}. \tag{1.1}$$

Tali quantità, essendo proporzioni, sono comprese tra 0 e 1.
Inoltre, essendo proporzioni, è possibile associare alla sensibilità e specificità un intervallo di confidenza di livello approssimato $(1-\alpha)$. La formula usualmente utilizzata per costruire tale intervallo per la sensibilità è

$$SN \pm z_{1-\alpha/2}\sqrt{\frac{SN(1-SN)}{n_{SN}}}, \tag{1.2}$$

con $z_{1-\alpha/2}$ quantile della distribuzione normale di livello $1-\alpha/2$ e $n_{SN} = FN + VP$. Ovviamente, la (1.2), con le opportune variazioni, può essere utilizzata anche per il calcolo dell'intervallo di confidenza per la specificità, sostituendo SN e n_{SN} con, rispettivamente, SP e n_{SP}. Si ottiene

$$SP \pm z_{1-\alpha/2}\sqrt{\frac{SP(1-SP)}{n_{SP}}}\,. \tag{1.3}$$

Analogamente, n_{SP} corrisponde alla somma di *VN* e *FP* e costituisce il totale dei pazienti sani.

In generale, indicando con y la realizzazione di una variabile binomiale di parametri n e p, la proporzione campionaria è $\hat{p} = \frac{y}{n}$. Il corrispondente **intervallo di confidenza alla Wald** (si veda, ad esempio, Pace e Salvan, 1996, Cap. 3), basato sulla normalità asintotica di $\hat{p}$, assume la forma

$$IC_W = \hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,. \tag{1.4}$$

Pur essendo l'intervallo di confidenza più utilizzato, l'intervallo alla Wald può risultare inaccurato per valori di p prossimi a 0 o a 1, o anche per valori moderati di n (si veda Agresti e Coull, 1998; Brown *et al.*, 2001).

Alcuni studi in letteratura (Volleset, 1993; Newcombe, 1998; Agresti e Coull, 1998; Brown *et al.*, 2001) hanno mostrato che vi sono almeno due intervalli che possono essere preferibili a (1.4).

Il primo di questi è l'**intervallo *score*** (si veda ad esempio Pace e Salvan, 2006, Cap. 3), che assume la forma

$$IC_S = \frac{\left(\hat{p} + \frac{z^2_{1-\alpha/2}}{2n} \pm z_{1-\alpha/2}\sqrt{\frac{\left(\hat{p}(1-\hat{p}) + z^2_{1-\alpha/2}/4n\right)}{n}}\right)}{\left(1 + \frac{z^2_{1-\alpha/2}}{n}\right)}\,. \tag{1.5}$$

Tale intervallo è noto anche come intervallo di Wilson (1927).

In alternativa, Agresti e Coull (1998) propongono un intervallo di confidenza estremamente semplice da calcolare, ma che presenta buone probabilità di copertura. Tale intervallo viene costituito con la semplice regola di "aggiungere due successi e due insuccessi" e quindi di applicare la formula usuale alla Wald.

Per tale motivo, questo intervallo viene detto **intervallo di confidenza di Wald aggiustato** e utilizza la formula (1.4), ma con numero di prove pari a $\tilde{n} = n + 4$ e stima di p data da $\tilde{p} = \frac{y+2}{\tilde{n}}$. L'intervallo è dato da

$$IC_{AC} = \tilde{p} \pm z_{1-\alpha/2} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}},$$

e risulta accurato, in particolare, in caso di numerosità campionarie piccole e per valori di p prossimi a 0 o a 1. Per valori piccoli della dimensione campionaria è preferibile l'intervallo di Wald aggiustato, mentre per valori grandi gli intervalli classici, *score* e di Wald aggiustati sono equivalenti.

In conclusione, gli intervalli *score* e gli intervalli di confidenza di Wald aggiustati forniscono intervalli con una probabilità di copertura attuale più vicina al livello di confidenza nominale. Inoltre, i risultati di simulazione ottenuti con questi sono migliori rispetto a quelli ottenuti con gli intervalli esatti (Agresti e Coull, 1998). Infatti, vengono spesso preferiti risultati approssimati rispetto a quelli esatti in quanto sono meno conservativi.

Si considerino ora due esperimenti indipendenti di conteggio binomiale, descritti da due variabili casuali indipendenti $Y_1 \sim Bi(n_1, p_1)$ e $Y_2 \sim Bi(n_2, p_2)$, con osservazioni y_1 e y_2, rispettivamente. Si perviene a tale modello partendo da $n = n_1 + n_2$ dati binari, classificati in due insiemi, relativi a popolazioni che possono essere non omogenee. Si assume che la probabilità di successo di una singola prova sia pari a p_1 se l'unità appartiene all'insieme 1 (per esempio, come nel caso d'interesse che un paziente scelto a caso risulta positivo ad un dato test). Sia invece pari a p_2 se appartiene all'insieme 2 (per esempio, rimanendo in tema, un paziente scelto casualmente risulta positivo, sottoposto però ad un altro esame). I dati da analizzare sono rappresentabili in una tabella di contingenza 2x2, come illustrato nella Tabella 1.1.

Il seguente rapporto

$$\frac{p_1}{1-p_1},$$

viene definito ***odds***, ossia il rapporto tra la probabilità di successo e la probabilità di insuccesso. Il logaritmo degli *odds* viene chiamato logit; ed è dato da

$$\mathrm{logit}(p_1) = \log\left(\frac{p_1}{1-p_1}\right).$$

Mentre il rapporto tra i rispettivi *odds* viene detto ***odds ratio***, ed è dato da

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)}.$$

Analogamente si può calcolarne la trasformazione logaritmica

$$\log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right) = \mathrm{logit}(p_1)\text{-}\mathrm{logit}(p_2).$$

Un valore dell'*odds ratio* pari a 1, implica che gli *odds* dell'evento, ossia il rapporto di successo contro insuccesso, sono uguali nei due gruppi, e quindi il verificarsi dell'evento è indipendente dalla variabile che distingue i due gruppi (nel nostro esempio il test applicato per valutare se un paziente è malato). Il *log-odds ratio* è, dunque, una misura di differenza tra gli *odds,* interpretabile in termini di confronto tra le probabilità p_1 e p_2.

E' spesso di interesse saggiare l'ipotesi nulla $H_0 : p_1 = p_2$, che esprime il fatto che la probabilità di successo è uguale nei due esperimenti. Tale aspetto sarà ripreso nel Capitolo 4.

Per delle trattazioni di questi argomenti si vedano: Pace e Salvan (1996, Cap. 6) e Piccolo (1998, Cap. 24).

1.3 Test diagnostici per l’individuazione dell’infezione da micobatteri

Per la diagnosi della tubercolosi possono essere eseguiti due tipi di approcci: diretto e indiretto. In questo studio si farà riferimento solo al primo approccio.

L’approccio <u>diretto</u> identifica il *Mycobacterium Tuberculosis* (MTB) mediante tre possibili tecniche:

A. Microscopia diretta;
B. Coltura;
C. PCR (*Polimerase Chain Reaction*, ovvero reazione a catena della polimerasi).

La **microscopia diretta** è rappresentata da due analisi: nella prima fase, quella esplorativa, si effettua un esame microscopico del linfonodo, cioè un’osservazione diretta con il vetrino e se ne osservano le caratteristiche e composizione. Nella fase successiva, si esegue sullo stesso linfonodo un esame microsopico in cui, però, esso viene sottoposto alla colorazione Ziehl-Neelsen. In sostanza, l’esecuzione della colorazione è suggerita dal fatto che il primo vetrino ha rilevato delle lesioni granulomatose e si sospetta la presenza di micobatteri. La Ziehl-Neelsen servirebbe, pertanto, come conferma.

Le caratteristiche della microscopia diretta sono:

- <u>analisi</u>: si analizzano i linfonodi con il microscopio e si valuta la presenza di lesioni granulomatose in via esplorativa e successivamente si verifica lo stato di colorazione di Ziehl-Neelsen (Z-N);
- <u>bontà del risultato</u>: non è un esame molto attendibile. Ha una specificità dell’ 80% e una sensibilità del 30%;
- <u>tempistica</u>: il risultato è immediato e può assumere i valori positivo o negativo.

Se il test risulta negativo, non è possibile escludere l’infezione.

Affinché sia possibile rilevare microscopicamente la presenza di bacilli acido-alcol resistenti, il materiale biologico in esame deve contenerne almeno 5-10x10^3 micobatteri per ml.
L’osservazione diretta del singolo campione ha una sensibilità che varia da 30% a 80% rispetto alla coltura, test spiegato di seguito, e dipende dal tipo di campione, dalla specie micobatterica, dalla popolazione che afferisce al laboratorio, dal metodo di rilevamento utilizzato e dall’esperienza di chi legge il preparato microscopico. L'esame microscopico è un elemento importante ai fini della valutazione della contagiosità del paziente, essendo questa direttamente correlata al numero di micobatteri presenti nelle secrezioni polmonari.

La **coltura** consiste in un’analisi sul linfonodo per verificare la presenza di micobatteri. La coltivazione dei batteri in laboratorio viene effettuata prelevando del materiale dalle lesioni che si suppone possano essere dovute al micobatterio. Successivamente, si coltivano questi batteri depositandoli su determinati “terreni di coltura”, cioè terreni o mezzi di coltura utilizzati per riprodurre artificialmente un ambiente in grado di soddisfare le esigenze del batterio che si desidera coltivare. Infine, si osserva il terreno e si verifica se la carica di micobatteri è cresciuta, valutandone anche la sistemazione. Dal risultato ottenuto si capisce il tipo di micobatterio che ha provocato quella lesione.
Le caratteristiche di questo esame sono:

- analisi: si tratta di un test sul linfonodo per verificare la presenza del MB;
- bontà del risultato: molto specifico e sensibile. E’ pertanto il metodo più attendibile per individuare la patologia;
- tempistica: dai 7-15 gg alle 3-6 settimane.

Per la coltura sono sufficienti da 10 a 100 micobatteri/ml. Il tempo medio per la coltura di un ceppo dei MTB si diversifica in base al tipo di terreno utilizzato: se

il terreno è liquido, è di circa 7-15 giorni; mentre le colture sui tradizionali terreni solidi necessitano in media di 3-6 settimane. Alcuni ceppi micobatterici crescono solo sui terreni solidi.
In conclusione, i tempi di refertazione di un esame colturale negativo sono:
- in terreno liquido, 6 settimane;
- in terreno solido, 8 settimane.

Tempi così lunghi per avere l'esito del test provocano un ritardo nella scelta dell'iter terapeutico più adeguato e aumentano il rischio di contagio.

La **PCR** è una tecnica diagnostica nuova e costituisce il nuovo test oggetto di studio. Questo esame consiste in una tecnica di biologia molecolare che si propone di amplificare il DNA estratto dai linfonodi dei pazienti per ottenere la reazione desiderata.
Le sue caratteristiche sono:

- <u>analisi</u>: test per l'amplificazione acidi nucleici, cioè una tecnica atta ad amplificare il DNA;
- <u>bontà del risultato</u>: si vuole testare la bontà di questa analisi per eventualmente utilizzarla in futuro in casi simili;
- <u>tempistica</u>: dipende dalla velocità con cui si trova la buona combinazione dei componenti e se la si scopre.

Se l'esito è negativo non è detto che non ci sia infezione.
I test di amplificazione degli acidi nucleici trovano indicazione solo nella fase diagnostica e non nel follow-up della TB. Essi permettono di rilevare la presenza di MTC nel materiale biologico entro poche ore dal prelievo del campione, ma non sostituiscono l'esame microscopico e l'esame colturale poiché amplificano il DNA o l'RNA ribosomiale di micobatteri sia vivi che morti.
L'esame microscopico e l'esame colturale devono invece essere eseguiti sempre per valutare l'infettività del paziente, confermare o meno la presenza di Micobatteri vitali e permettere l'allestimento delle prove di farmacosensibilità "in

vitro". Non si deve utilizzare il materiale biologico per l'esecuzione dei test di amplificazione se questo compromette la possibilità di eseguire l'esame microscopico e l'esame colturale.

Attualmente la PCR non è un esame di routine e trova indicazione:

a) nei casi di esame microscopico positivo per anticipare l'identificazione;
b) in presenza di forte sospetto clinico nonostante la negatività dell'esame microscopico, per aumentare la probabilità di diagnosi.

In sintesi, un confronto diretto tra i tre approcci viene sintetizzato nella Tabella 1.2.

		Tecniche diagnostiche		
Caratteristiche		**Microscopia diretta**	**Coltura**	**PCR**
	Quantità materiale	5-10 X 10^3 MB/ml	10-100 MB/ml	1 molecola di DNA
	Tempo impiegato	immediato	da 7 gg a 6 settimane	dipende dalla combinazione
	Bontà del risultato	inaffidabile	affidabile	*da testare*

Tabella 1.2: *Confronto tra test diagnostici per la TB.*

1.4 I dati e le variabili del caso di studio di Padova

Il dataset considerato in questa tesi è composto da pazienti estratti in base a referti ricavati dal database "Armonia" dell'Istituto di Anatomia Patologica e, poi, analizzando le cartelle cliniche dei medesimi pazienti custodite presso l'archivio dell'Azienda Ospedaliera di Padova.

Lo studio prende in esame 28 pazienti adulti ricoverati tra il 2002 e il 2007. A questi pazienti sono stati asportati linfonodi il cui esame microscopico diretto presenta delle lesioni granulomatose sospette per un'infezione micobatterica.
I linfonodi di ciascun paziente, dopo l'estrazione, sono stati inclusi in paraffina e, così conservati, sono stati utilizzati in questo studio allo scopo di sperimentare la PCR nell'identificazione dell'infezione da tubercolosi.
Per ogni linfonodo sono stati svolti i seguenti test:

- esame <u>microscopico</u> diretto e successiva colorazione Z-N;
- esame <u>colturale</u>, che è stato recuperato dall'archivio informatico del Dipartimento di Microbiologia e/o dalla cartella clinica del paziente, in quanto è un'analisi che si può eseguire solo con linfonodi freschi;
- esame della <u>PCR</u>.

Il protocollo per eseguire la PCR richiede:

1. di tagliare i linfonodi in piccole parti;
2. la de-paraffinizzazione del linfonodo, che consiste in lavaggi nello xilolo ed etanolo con successiva essiccazione e risospensione del materiale ottenuto;
3. l'estrazione del DNA;
4. il controllo interno, cioè la **PCR per la beta-globina** o beta actina. E' un controllo utilizzato per verificare la qualità del DNA estratto, verificando, inoltre, se è possibile proseguire con la PCR specifica per micobatteri su tale estratto (quindi se il materiale estratto è sufficiente e di buona qualità);
5. la **PCR per MB**: sui campioni per i quali la PCR per beta globina è riuscita, si può procedere con la PCR per MB.

Per effettuare il test PCR sono state utilizzate due tecniche. La prima (Tecnica A) utilizza un estrattore automatico (utilizzato prevalentemente dagli analisti di laboratorio), mentre la seconda (Tecnica B) è più manuale e utilizza più particelle del campione per l'analisi. La Tecnica A è più aggressiva rispetto alla B in fase di lavaggio. Le principali differenze nelle due teniche sono riportate nella Tabella 1.3.

Tecniche	**A**	**B**
Parti di linfonodo utilizzate	5	7
De-paraffinizzazione	2 lavaggi con xilene e 2 lavaggi con etanolo	1 lavaggio con xilene e 2 lavaggi con etanolo
Estrazione DNA	estrattore automatico	manuale

Tabella 1.3: *Confronto tra la Tecnica A e la Tecnica B per effettuare il test PCR.*

Nel dataset, per ogni paziente sono disponibili dati personali, quali genere ed età al momento del ricovero. Inoltre nel dataset sono presenti i risultati delle tecniche applicate ai linfonodi. In particolare si rileva: microscopia diretta con e senza colorazione Z-N (vetrino), coltura, controllo interno sia per la prima che per la seconda tecnica e PCR con entrambe le tecniche. Le variabili sono elencate nella Tabella 1.4.

NOMI	DEFINIZIONE	COMMENTO
Nome	Indica il nominativo del paziente ricoverato	
Genere	Rappresenta il genere dei pazienti	Dicotomica (M o F)
Data nascita	Data di nascita del paziente	
Data ricovero	Data del ricovero in reparto del paziente	
Età	Identifica l'età del paziente al momento del ricovero	Assume valori da 20 a 82
Reparto	Indica il reparto in cui il paziente è stato ricoverato	
Vetrino	Rappresenta la descrizione del linfonodo osservato attraverso il mocroscopio	Qualitativa con livelli 1, 2 e 3 che descrivono lo stato granulomatoso del linfonodo
Colorazione ZN	Analisi del linfonodo con microscopio dopo averlo sottoposto alla colorazione Ziehl-Neelsen	Dicotomica: assume valore positivo (presenza di MB) o negativo (assenza di MB)
PCR beta globina A	Controllo della qualità del linfonodo con estrazione del DNA per il gruppo A	Qualitativa con i livelli: assente, non estratta e positiva **(*)**
PCR beta globina B	Controllo della qualità del linfonodo con estrazione del DNA per il gruppo B	Qualitativa che assume valore positiva e non estratta
PCR MB A	Esame della PCR per il campione della Tecnica A	Qualitativa e può essere classificata in positiva, fallita ed assente
PCR MB B	Esame della PCR per il campione della Tecnica B	Qualitativa e può essere classificata in positiva, fallita ed assente
Colture	Esito del test coltura	Qualitativa con livello positivo, negativo o assente

(*) la categoria assente comprende un unico caso, nel quale lo strumento utilizzato per la Tecnica A ha, per errore, perso un linfondo e quindi non si è riusciti ad eseguire i controlli.

Tabella 1.4: *Riassunto delle variabili presenti nel dataset.*

1.5 Considerazioni conclusive

Questo capitolo, oltre ad introdurre il problema e il caso di studio, fornisce una prima descrizione del dataset e delle sue variabili.
Nel Capitolo 2 sarà illustrata una prima analisi delle variabili del dataset. Queste analisi sono utili, in primo luogo, per fornire un’introduzione alla struttura del dataset e, in secondo luogo, per individuare eventuali legami e interazioni tra le variabili. Inoltre, il Capitolo 2 contiene i primi commenti e analisi per verificare se la nuova tecnica in via di sperimentazione, la PCR, ottiene risultati attendibili e importanti per una sua eventuale applicazione in fase di formulazione della diagnosi al fine di individuarne la patologia.

Capitolo 2

Una preliminare analisi del campione

In questo capitolo si presenta con più dettaglio il dataset introdotto nel capitolo precedente e si illustrano le caratteristiche specifiche del campione preso in esame.

Si ricorda che i dati sono stati forniti dal Professor Fassina attraverso un'attenta selezione tra i database dell'Azienda Ospedaliera di Padova e che lo scopo principale di questa analisi è fornire uno strumento valido per la valutazione dei test diagnostici.

I dati sono interpretabili solo dopo aver calcolato quantità che riassumono le caratteristiche salienti delle variabili di interesse. Nel seguito, si considerano alcune semplici tecniche di sintesi numerica e grafica che si applicano a singole variabili o a coppie di variabili. Inoltre, sono utilizzate anche alcune tecniche di base di inferenza statistica.

Per la stesura di questo capitolo si è fatto, principalmente, riferimento per la parte teorica ai testi di Pace e Salvan (1996) e Piccolo (2000); mentre per la parte applicativa si vedano i testi di Bortot *et al.* (2000) e di Iacus e Masarotto (2003).

2.1 Le variabili

Questa analisi ha lo scopo di individuare la struttura, composizione e la natura delle variabili principali del dataset. Sono state scartate alcune varabili che, data la loro natura, non sono utili nell'analisi, ossia quelle variabili che non forniscono informazioni aggiuntive per lo studio del micobatterio, come ad esempio la variabile nome, data di nascita e di ricovero, reparto, ecc.
La distribuzione di frequenza della variabile **Genere**, rilevata sui 28 pazienti, è riportata nella Tabella 2.1. Si può notare che tra i pazienti si ha il 54% di femmine (F) contro il 46% di maschi (M). Il rapporto maschi contro femmine è $\frac{M}{F} = 0.87$.

GENERE	FREQUENZE ASSOLUTE	FREQUENZE PERCENTUALI
FEMMINE	15	53.57
MASCHI	13	46.43
TOTALE	28	100

Tabella 2.1: *Valori assoluti e percentuali della variabile* ***Genere****.*

La distribuzione della variabile **Età** dei pazienti è rappresentata nella Figura 2.1.

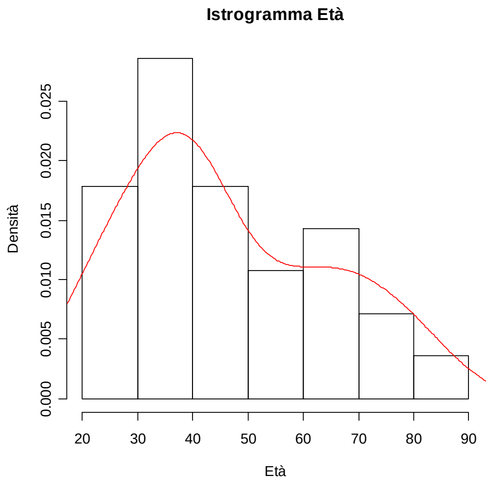
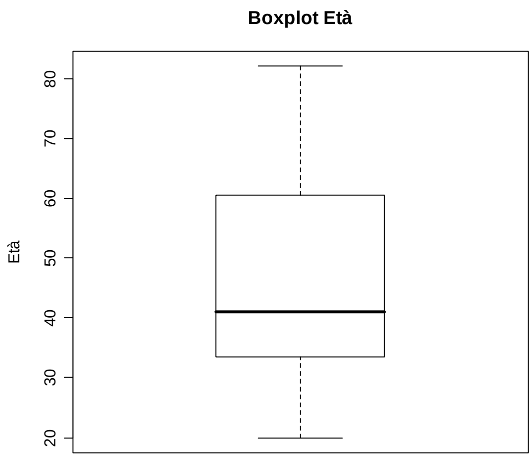


Figura 2.1: *Istogramma e densità secondo il metodo del nucleo della variabile **Età**.*

L'età media dei pazienti è di 46 anni (± 18.24), quella mediana di 41 e il dominio varia da 20 a 82 anni (si vedano la Figura 2.2 e la Tabella 2.2).



MINIMO	20
PRIMO QUARTILE	33.75
MEDIANA	41
MEDIA	46.04
TERZO QUARTILE	59.25
MASSIMO	82
ERRORE STANDARD	18.24

Figura 2.2 e Tabella 2.2: *Boxplot e valori di sintesi della variabile **Età**.*

Un test grafico molto utilizzato per verificare la provenienza di un insieme di dati da una popolazione normale (Figura 2.3) è il *q-q plot*. Dal diagramma q-q normale della variabile **Età** si nota che l'ipotesi di normalità può essere accettata. Anche il test di Shapiro per la normalità, che assume valore 0.93 (*p-value* = 0.06), o il test di Kolmogorov-Smirnov (valore = 0.18, *p-value*=0.32) portano ad accettare l'ipotesi nulla di normalità della variabile **Età.**

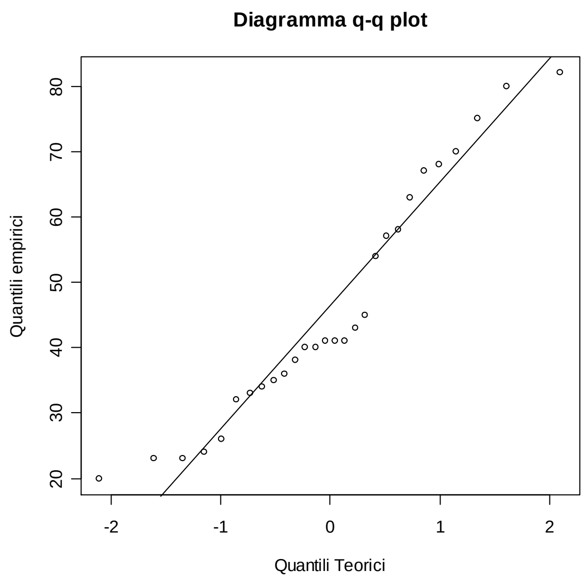


Figura 2.3: *Diagramma q-q normale per la variabile **Età**.*

La variabile **Vetrino** rappresenta i risultati ottenuti analizzando il linfonodo conservato in paraffina attraverso il microscopio. I valori che tale variabile può assumere sono stati assegnati come etichetta per identificare lo stato del linfonodo. In particolare si ha:

- 1=identifica flogosi (infiammazione) cronica granulomatosa gigantocell con aree necrosi[7];
- 2=rappresenta flogosi cronica granulomatosa gigantocell senza necrosi;
- 3=indica pseudotumore infiammatorio del linfonodo.

[7] *Necrosi*: processo irreversibile di morte delle cellule dei tessuti.

Le classi di questa variabile indicano che è presente una lesione granulomatosa nel linfonodo e quindi si sospetta un'infezione da micobatterio. Questa è la variabile utilizzata per selezionare i pazienti inseriti nello studio. In particolare in letteratura la più suggestiva delle classi è la prima, ma le altre devono comunque essere accertate con test aggiuntivi. E' pertanto una variabile relativamente poco importante nello studio, in quanto i valori si concentrano tutti nella prima classe (vedi Tabella 2.3).

VETRINO	**FREQUENZE ASSOLUTE**	**FREQUENZE PERCENTUALI**
1	25	89.29
2	2	7.14
3	1	3.57
TOTALE	28	100

Tabella 2.3 *Valori assoluti e percentuali della variabile **Vetrino**.*

La variabile **Colorazione ZN** rappresenta il risultato ottenuto dopo aver sottoposto il linfonodo alla colorazione Ziehl-Neelsen (Z-N). L'esame con il microscopio viene classificato in:

- negativo: l'esame non individua la presenza di MB;
- positivo: l'esame evidenzia la presenza di MB.

COLORAZIONE ZN	**FREQUENZE ASSOLUTE**	**FREQUENZE PERCENTUALI**
NEGATIVO	24	85.71
POSITIVO	4	14.29
TOTALE	28	100

Tabella 2.4 *Valori assoluti e percentuali della variabile che identifica la **Colorazione ZN**.*

Notiamo che per questa variabile ci sono solo il 14.29% dei risultati positivi e tutti i restanti indicano l'assenza di MB; tali conclusioni sono contrastanti rispetto a quelle ottenute per la variabile **Vetrino**.

La **PCR beta globina A** e **PCR beta globina B** sono entrambe variabili che misurano la qualità del DNA estratto, rispettivamente, per il gruppo A e B. In particolare, si analizza se la qualità dell'estrazione è sufficiente per l'analisi successiva, ossia l'esame PCR su micobatterio. Tali variabili possono assumere i seguenti valori:

- assente (solo per il gruppo A): lo strumento utilizzato con la Tecnica A ha erroneamente perso il linfonodo e perciò non è stato possibile svolgere il controllo;
- non estratta: il materiale è insufficiente e non è riuscita l'estrazione del DNA;
- positiva: il DNA estratto è conforme alle aspettative per poter applicare la PCR su MB.

La distribuzione di frequenza per le variabili **PCR beta globina A** e **PCR beta globina B** sono riportate nella Tabella 2.5.

	PCR BETA GLOBINA A		PCR BETA GLOBINA B	
	FREQ. ASSOLUTE	FREQ. %	FREQ. ASSOLUTE	FREQ. %
ASSENTE	1	3,57	0	0,00
NON ESTRATTA	9	32,14	6	21,43
POSITIVA	18	64,29	22	78,57
TOTALE	28	100	28	100

Tabella 2.5: *Distribuzioni di frequenza per la variabile **PCR per Beta Globina** in entrambi i gruppi.*

Come si vede dalla Tabella 2.5 la maggior parte dei campioni sono risultati positivi al test: il 64% per la Tecnica A e il 78 % per quella B. Per questa parte dei dati si proseguirà con l'esame della PCR su MB. Invece, i campioni per cui non è riuscita l'estrazione (compreso l'unico caso in cui è assente) faranno parte della macrocategoria "assente" della variabile **PCR MB** e per questi casi lo studio si interrompe.

L'esame successivamente eseguito è la PCR su Micobatterio, applicata ai campioni del gruppo A e del gruppo B. Questo esame viene identificato con le variabili: **PCR MB A** e **PCR MB B**. Le modalità assunte da tali variabili sono:

- <u>assente</u>: valori per cui il controllo precedente (**PCR beta globina**) è assente o non è riuscito per DNA insufficiente. In sostanza dal campione A originale sono stati decurtati 10 linfonodi: 9 perchè il controllo precedente non è riuscito e 1 che è andato perso durante l'esame di beta globina. Mentre per il gruppo B sono stati esclusi 6 linfonodi, in quanto per tali casi non è riuscita l'estrazione del DNA e sarebbe inutile procedere con il test;
- <u>negativa</u>: casi per cui non è riuscito l'esame PCR con la Tecnica A e B. Comprende, anche, eventuali casi in cui l'esame PCR è riuscito, in quanto si è concluso con successo, ma il risultato non ha trovato la presenza di un micobatterio. In particolare, un elemento del campione A risulta essere un microrganismo detto *saccharopolyspora erythraea*; per i campioni B vi sono due casi che risultano essere microrganismi detti *nocadie*, (probabilmente i pazienti relativi erano soggetti a un'infezione dovuta a quel microbatterio);
- <u>positivo</u>: l'esame è riuscito e il risultato ha dimostrato la presenza di micobatteri. Nei due casi in questione per la Tecnica A sono stati rilevati: un *M. Tuberolosis* e un micobatterio detto *kumamotonense*; mentre per la Tecnica B non si sono rilevati esiti positivi.

Le distribuzioni di frequenza delle variabili **PCR MB A** e **B** sono riportate nella Tabella 2.6.

	PCR MB A		**PCR MB B**	
	FREQ. ASSOLUTE	**FREQ. %**	**FREQ. ASSOLUTE**	**FREQ. %**
ASSENTE	10	35.71	6	21.43
NEGATIVA	16	57.14	22	78.57
POSITIVA	2	7.14	0	0.00
TOTALE	28	100	22	100

Tabella 2.6 *Distribuzioni di frequenza per le variabili* ***PCR su MB*** *per entrambe le tecniche.*

Dalla Tabella 2.6 emerge che la tecnica A non è molto efficiente, in quanto fallisce in ben 16 esperimenti (57.14%) e risulta positiva solo in 2 (7.14%). Il campione B, invece, appare peggiore del precedente perché non ci sono stati risultati positivi e tutti gli esiti sono stati negativi (anche se le unità analizzate sono state in quantità superiore rispetto al campione A).

L'ultima variabile del dataset, prende il nome **Coltura** e indica il risultato ottenuto con l'esame coltura effettuato al momento del ricovero. I valori assunti da tale variabile sono:

- assente: risultato del test non disponibile;
- negativo: risultato negativo alla presenza di micobatteri nel linfonodo;
- positivo: risultato positivo alla presenza di MB. Il paziente probabilmente era soggetto ad un'infezione degli stessi.

Questa variabile gioca un ruolo fondamentale nella valutazione della nuova tecnica, la PCR, e nel confrontare la bontà delle due tecniche sopra citate.
La distribuzione di frequenza della variabile **Coltura** è riportata nella Tabella 2.7.

COLTURA	FREQUENZE ASSOLUTE	FREQUENZE PERCENTUALI
ASSENTE	10	35.71
NEGATIVA	13	46.43
POSITIVA	5	17.86
TOTALE	28	100

Tabella 2.7: *Distribuzione di frequenza per la variabile* ***Coltura****.*

Da questa prima analisi notiamo che anche la variabile **Coltura** presenta solo 5 risultati positivi a differenza delle nostre aspettative. Dobbiamo tenere presente, però, che circa il 36% dei dati non sono stati recuperati e questo incide in maniera considerevole, vista la numerosità esigua del campione. Quest'ultima informazione influenzerà sicuramente la precisione delle analisi successive, in quanto è molto rilevante rispetto al totale.

In conclusione, le due tecniche sperimentali non sembrano essere adeguate per l'identificazione dei MB. In particolare, la tecnica A sembra migliore della B, ma comunque non abbastanza affidabile per rappresentare un test che anticipi l'esito della coltura.

Per un'analisi più accurata, nel paragrafo successivo, verranno confrontate le relazioni generali tra le variabili più importanti per individuare un'eventuale dipendenza tra queste. In particolare si accentua l'attenzione tra **PCR MB A – B,** e **Colorazione ZN** con **Coltura,** per valutare se gli effetti rilevati da queste corrispondono agli esiti reali della variabile **Coltura**.

2.2 Relazioni tra variabili

Nel paragrafo precedente ci siamo occupati dello studio delle distribuzioni delle variabili presenti nel dataset. Ne abbiamo studiato alcuni valori caratteristici e le rappresentazioni in tabella.

Il passo successivo e naturale è quello di vedere se esistono legami tra le coppie di variabili rilevate sui pazienti. In questo paragrafo si presentano pertanto i risultati di alcune analisi bivariate svolte tra le variabili introdotte nel paragrafo precedente. Viene esclusa da queste analisi la variabile **Vetrino,** in quanto utilizzata esclusivamente per selezionare i pazienti per lo studio.

Nel seguito, a seconda della natura delle variabili considerate, si farà riferimento ad alcune procedure di inferenza classiche: il test esatto di Fisher; il test *t* di Student per la verifica di ipotesi tra le medie di due campioni (vedi Pace e Salvan, 2001, Cap. 0 e 10).

Nella Tabella 2.8 sono riportati i *p-values* del test esatto di Fisher per verificare l'indipendenza tra le variabili. Si nota che tra le variabili **PCR beta globina A** e **PCR MB A** c'è un legame di dipendenza, considerazione ovvia visto che la prima rappresenta una selezione che influenza i risultati ottenuti nella seconda. Anche tra la **PCR beta globina B** e la **PCR MB B** vi è dipendenza, e anche per queste vale il commento precedente. Infatti, il test esatto di Fisher relativo a queste variabili, rifiuta l'ipotesi nulla di indipendenza. Le altre variabili risultano indipendenti tra loro. Nemmeno questo risultato è un'indicazione positiva per il nuovo test diagnostico. Infatti, in realtà ci aspettavamo una dipendenza maggiore tra le variabili d'interesse **PCR MB** e **Coltura**.

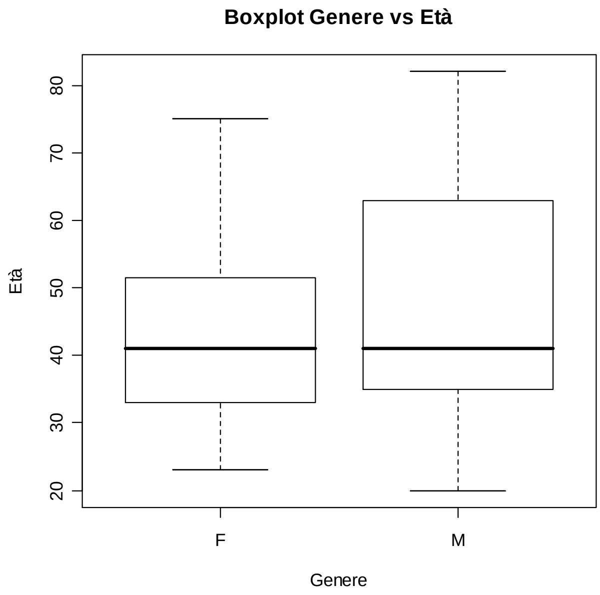
Test Esatto Di Fisher (p_value)	**Genere**	**Colorazione ZN**	**PCR beta globina A**	**PCR beta globina B**	**PCR MB A**	**PCR MB B**	**Colture**
Genere		1.00	0.68	0.65	0.47	0.65	1.00
Colorazione ZN			0.27	1.00	0.41	1.00	0.49
PCR beta globina A				1.00	4.89e-07 (*)	1.00	0.6
PCR beta globina B					1.00	2.65e-06 (*)	1.00
PCR MB A						1.00	0.37
PCR MB B							1.00
Colture							

(*) Indica *p-value* < 0.05

Tabella 2.8 *P-value del test esatto di Fisher.*

Nel seguito si confrontano le medie di due gruppi o due popolazioni. Consideriamo, inizialmente, l'effetto del **Genere** su alcune variabili rilevate nel dataset.

In via esplorativa, per la variabile **Età** tale relazione viene rappresentata nella Figura 2.4, in cui viene riportato il boxplot della variabile **Età** rispetto al **Genere**. Si nota che la media dei due gruppi si mantiene attorno a 40 anni. Le pazienti sono più concentrate nella fascia d'età dai 33 ai 50 anni circa, mentre l'età dei maschi è meno concentrata. La Tabella 2.9 riporta, invece, la numerosità dei due gruppi, le medie e gli errori standard.



	M	F
Numerosità	13	15
Media	48.77	43.67
Errori Standard	20.27	16.62

Figura 2.4 e Tabella 2.9: *Boxplot della variabile* ***Età*** *rispetto al* ***Genere*** *e valori riassuntivi suddivisi per* ***Genere****.*

Per valutare la normalità della variabile **Età** si considera il diagramma q-q normale (Figura 2.5) e si esegue il test di Shapiro (*p-value*= 0.52 per i maschi e *p-value* = 0.12 per le femmine).

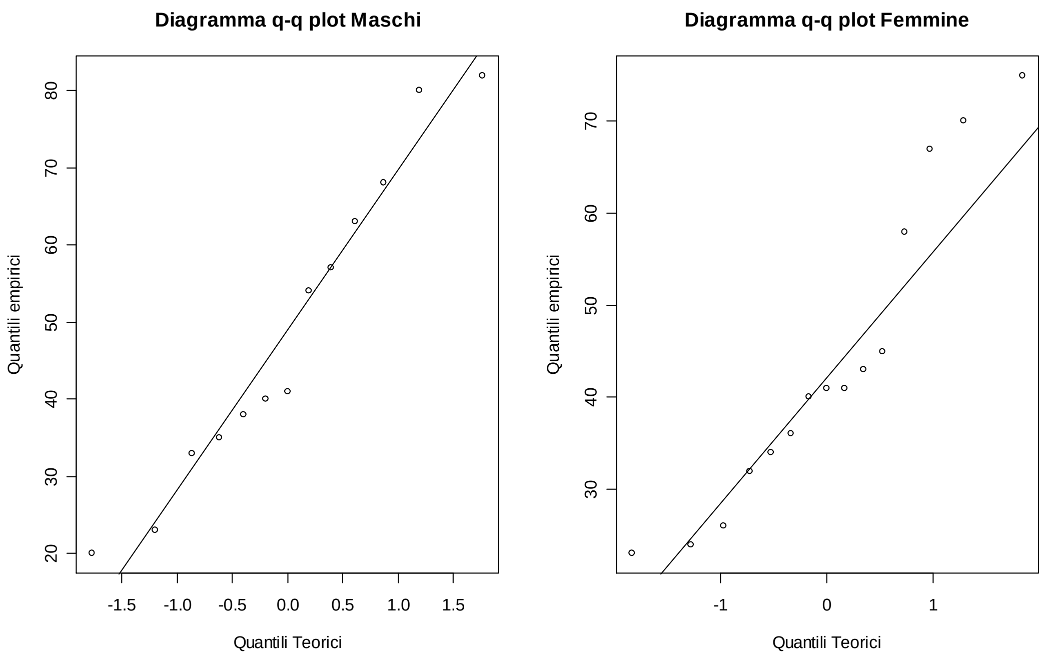


Figura 2.5: *Diagramma q-q normali della variabile* ***Età*** *suddivisi per* ***Genere****.*

Per entrambi i diagrammi q-q normali l'ipotesi di normalità può essere accettata. Anche i risultati del test di Shapiro confermano tale considerazione; quindi, si accetta l'ipotesi nulla di normalità di entrambi i gruppi.

Se si vuole valutare la differenza tra le medie nei due gruppi, è necessario eseguire, per verificare l'omoschedasticità delle variabili, il test F (Pace e Salvan, Cap. 7). Il valore della statistica è 1.49 e il relativo *p-value* è 0.47. Si accetta pertanto l'ipotesi nulla di uguaglianza delle varianze delle variabili nei due gruppi. Verificata la normalità e omoschedasticità, si procede con lo studio dell'uguaglianza delle medie con il test *t* di Student. Il valore del test è 0.72, con 23 gradi di libertà, e *p-value* pari a 0.48. Si accetta, pertanto, l'ipotesi nulla di uguaglianza delle medie.

Si considerano ora gli incroci tra le variabili oggetto di studio, ovvero **PCR MB A**, **PCR MB B** e variabile **Coltura**, che certifica l'effettiva presenza del micobatterio nel linfonodo del paziente, ossia che identifica l'appartenenza alla categoria sani/malati dei pazienti.

		COLTURA	
PCR MB A		**NEGATIVA**	**POSITIVA**
	NEGATIVA	7	3
	POSITIVA	0	1

Tabella 2.10: *Relazione tra la nuova tecnica* ***PCR MB A*** *e l'esame classico* ***Coltura***.

I risulati della Tabella 2.10 non sembrano molto incoraggianti. La **Coltura** individua 4 esiti positivi al test (36%), mentre la nuova tecnica **PCR MB A** ne individua correttamente solo 1 (9%). Quindi i risultati concordano in 8 casi, 73% (I.C.: 0.47, 0.99). Queste considerazioni sembrano dimostrare che la Tecnica A non è molto indicata per questo tipo di analisi e non sembra essere efficiente come esame preliminare.

		COLTURA	
PCR MB B		**NEGATIVA**	**POSITIVA**
	NEGATIVA	10	4
	POSITIVA	0	0

Tabella 2.11: *Relazione tra la nuova tecnica* ***PCR MB B*** *e l'esame classico* ***Coltura***.

Confrontando invece la **Coltura** con la **PCR MB B,** come illustrato dalla Tabella 2.11, notiamo che la **PCR MB B** non rileva esiti positivi e i risultati concordano per il 71% (I.C.: 0.47, 0.95) con la **Coltura**. Questo incrocio dimostra che la Tecnica B non riesce ad identificare correttamente i pazienti che realmente manifestano la patologia.

Analizziamo, ora, la relazione tra le variabili **PCR MB** per i due gruppi nella Tabella 2.12.

		PCR MB B	
PCR MB A		**NEGATIVA**	**POSITIVA**
	NEGATIVA	12	0
	POSITIVA	2	0

Tabella 2.12: *Relazione tra le nuove tecniche* ***PCR MB A*** *e* ***B****.*

Vediamo che i risultati concordanti sono 12, 86% dei pazienti (I.C.: 0.68, 1.00). Dal confronto fra queste due Tecniche, ci aspettavamo una concordanza maggiore negli esiti, infatti, sono test eseguiti sui medesimi campioni. Invece, come si può notare dalla Tabella 2.12, non sembrano coincidere molto.
Tra le due tecniche, la **PCR MB B** come notato prima, non ha risultati positivi, mentre il gruppo A ne ha due. Pertanto, quest'ultima sembra essere la migliore tra le due.

Vediamo, ora la relazione tra la variabile **Colorazione ZN** e **Coltura** nella Tabella 2.13.

		COLTURA	
Colorazione ZN		**NEGATIVA**	**POSITIVA**
	NEGATIVA	12	4
	POSITIVA	1	1

Tabella 2.13: *Relazione tra la variabile* ***Colorazione ZN*** *e* ***Coltura****.*

Dall'incrocio si rilevano il 72% (I.C.: 0.51, 0.93) dei risultati concordanti, un tasso abbastanza alto rispetto alle analisi precedenti. Questa tecnica rappresenta una buona analisi preliminare anche se i risultati positivi vengono rilevati difficilmente come, anche, per le due tecniche precedenti.

2.3 Considerazioni conclusive

In questo capitolo, e in particolare in queste ultime pagine, sono stati presentati i risultati di alcune analisi statistiche di base. Queste procedure di inferenza fanno riferimento alla teoria standard della verosimiglianza e alle approssimazioni, semplici e generali, per le distribuzioni campionarie delle quantità di verosimiglianza. Tali risultati vengono detti del primo ordine per distinguerli dai più avanzati risultati di ordine superiore di cui ci si occuperà nel prossimo capitolo.

Nel prossimo capitolo sono infatti presentate alcune recenti tecniche asintotiche per lo studio di campioni con numerosità limitata. Queste consistono in miglioramenti dei test classici che permettono l'aumento della potenza dei test, nonostante una dimensione campionaria esigua.

Inoltre, vengono applicate le nuove teorie al caso oggetto di studio per individuare i risvolti operativi che possono produrre conclusioni più affidabili anche in ambito medico.

Capitolo 3

Verosimiglianza e metodi asintotici

Obiettivo di questo capitolo è richiamare sinteticamente alcuni metodi inferenziali della teoria classica e moderna centrata sulla verosimiglianza. Alcune procedure di inferenza classiche sono state applicate al dataset oggetto di studio nel capitolo precedente. Tecniche più recenti saranno invece investigate per l'elaborazione dei dati nel Capitolo 4.

In primo luogo saranno richiamati alcuni risultati e le tecniche elementari dell'inferenza basata sulla funzione di verosimiglianza. Saranno riprese, in particolare, alcune quantità di verosimiglianza e la corrispondente teoria asintotica del primo ordine, introducendo le notazioni utilizzate nel resto della tesi. Tali quantità saranno illustrate, con particolare riferimento, alle famiglie esponenziali. Per questi argomenti, alcuni riferimenti sono, ad esempio, Azzalini (1992), Pace e Salvan (1996, Capp.1-3 e 5) e Severini (2000, Capp. 1, 3 e 4).

In secondo luogo saranno presentati alcuni recenti risultati relativi ai metodi asintotici di ordine superiore per quantità di verosimiglianza, quali il test del log-rapporto di verosimiglianza (si veda Pace e Salvan, 1996, Capp. 4 e 11; Severini, 2000, Capp. 5, 7 e 9; Brazzale *et al*., 2007, Cap. 8). Queste tecniche migliorano le approssimazioni del primo ordine, in particolare in presenza di parametri di disturbo di dimensione elevata e/o in caso di numerosità campionaria esigua.

3.1. La verosimiglianza

Si assuma che i dati siano costituiti da n osservazioni $y = (y_1, \ldots, y_n)$ indipendenti e identicamente distribuite (i.i.d.), realizzazioni di una variabile casuale (v.c.) Y con funzione di densità $f(y;\theta)$ indicizzata dal parametro θ, con $\theta \in \Theta \subseteq R^p$, $p \geq 1$.

La **funzione di verosimiglianza** per θ è

$$L(\theta) = L(\theta; y) = c(y)\prod_{i=1}^{n} f(y_i;\theta), \qquad (3.1)$$

con $c(y) > 0$ costante di proporzionalità arbitraria, indipendente da θ. Poichè $L(\theta)$ è non negativa, spesso può risultare conveniente considerare in luogo della funzione di verosimiglianza il suo logaritmo naturale, ovvero la **funzione di log-verosimiglianza**

$$l(\theta) = \log L(\theta; y) = c^*(y) + \sum_{i=1}^{n} \log f(y_i;\theta), \qquad (3.2)$$

con $c^*(y) = \log c(y)$ costante additiva arbitraria.

Nel seguito si assume che la log-verosimiglianza (3.2) sia una funzione di θ sufficientemente regolare, ossia che essa ammetta derivate parziali fino agli ordini richiesti. Inoltre, in generale, quantità definite a partire dalla funzione di verosimiglianza saranno dette quantità di verosimiglianza.

Le derivate di $l(\theta)$ fino al secondo ordine rivestono un ruolo centrale per l'inferenza; le derivate successive sono invece importanti per raffinamenti della teoria asintotica dell'inferenza.

Tra le quantità di verosimiglianza più importanti troviamo:

1) la **funzione punteggio** (*score*) di verosimiglianza, data da

$$l_*(\theta) = \frac{\partial l(\theta)}{\partial \theta}\,; \tag{3.3}$$

2) la matrice di **informazione osservata di Fisher**, data dalla matrice delle derivate parziali seconde di $l(\theta)$ cambiate di segno, ossia

$$j(\theta) \; = -l_{**}(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta\, \partial \theta^T}\,; \tag{3.4}$$

3) la matrice di **informazione attesa** di Fisher, data dal valore atteso dell'informazione osservata, ossia

$$i(\theta) \; = E_\theta(j(\theta)) = E_\theta\left(-\frac{\partial^2 l(\theta)}{\partial \theta\, \partial \theta^T}\right), \tag{3.5}$$

dove $E_\theta(\cdot)$ denota il valore atteso calcolato rispetto a $f(y;\theta)$;

4) qualora $l(\theta)$ sia differenziabile, il valore di θ tale per cui

$$l_*(\theta) = 0 \tag{3.6}$$

viene detto **stima di massima verosimiglianza** (s.m.v.) di θ ed è indicato con $\hat{\theta}$. Si assume nel seguito che la s.m.v. sia unica e che sia soluzione dell'equazione di verosimiglianza (3.6).

<u>Esempio 3.1</u>: Se le osservazioni sono tratte da una v.c. con densità appartenente ad una famiglia esponenziale naturale, allora la densità congiunta assume la forma

$$f(y;\theta) = h(y)\exp\{\theta^T t - K(\theta)\}, \tag{3.7}$$

dove $t = t(y)$ indica la statistica sufficiente minimale, θ è il parametro naturale, e $h(\cdot)$ e $K(\cdot)$ sono funzioni note.

La log-verosimiglianza è

$$l(\theta) = l(\theta; y) = \theta^T t - K(\theta)\,. \tag{3.8}$$

Il vettore *score* è

$$l_*(\theta) = t - \frac{\partial K(\theta)}{\partial \theta} = t - E_\theta(T)\,, \qquad (3.9)$$

con $E_\theta(T) = \frac{\partial K(\theta)}{\partial \theta}$ vettore delle medie. La (3.9) afferma che $l_*(\theta)$ coincide con il vettore degli scarti di t dal proprio valore atteso.

L'informazione osservata di Fisher è

$$j(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta\, \partial \theta^T} = \frac{\partial^2 K(\theta)}{\partial \theta\, \partial \theta^T} = Var_\theta(T) \qquad (3.10)$$

e, poiché $j(\theta)$ non dipende da t, essa coincide con l'informazione attesa $i(\theta)$.

In una famiglia esponenziale, se esiste, la s.m.v. di θ, è data dalla soluzione dell'equazione di verosimiglianza

$$t - \frac{\partial K(\theta)}{\partial \theta} = t - E_\theta(T) = 0\,. \qquad (3.11)$$

Se esiste, tale soluzione è unica poiché la matrice

$$-\frac{\partial^2}{\partial \theta\, \partial \theta^T} l(\theta) = Var_\theta(T) \qquad (3.12)$$

è definita positiva.

3.2. Teoria asintotica del primo ordine

La funzione di verosimiglianza e le quantità ad essa collegate costituiscono un riassunto dell'informazione contenuta nei dati e nel modello statistico adottato, utile per localizzare in modo naturale il modello probabilistico generatore dei dati. Procedure di verosimiglianza esistono sotto ipotesi tenui e, con ampia generalità, esse sono semplici e accurate.

Uno dei motivi principali per il successo dei metodi basati sulla verosimiglianza è la possibilità di ricorrere ad approssimazioni semplici e generali per le corrispondenti distribuzioni campionarie. Tali approssimazioni sono basate su approssimazioni limite della Teoria della Probabilità, che forniscono risultati validi al divergere della numerosità campionaria, ossia per $n \to +\infty$.

Nel seguito, si richiamano alcuni ben noti risultati di convergenza in distribuzione per le usuali quantità di verosimiglianza, che sono detti **del primo ordine** (vedi ad esempio Pace e Salvan, 1996, § 3.5). Si considerano le quantità di verosimiglianza *score* $l_*(\theta)$, il log-rapporto di verosimiglianza $W(\theta) = 2\left(l(\hat{\theta}) - l(\theta)\right)$, il test *score* $W_u = W_u(\theta) = l_*(\theta)^T i(\theta)^{-1} l_*(\theta)$ e il test Wald $W_e = W_e(\theta) = (\hat{\theta} - \theta)^T i(\theta)(\hat{\theta} - \theta)$.

I test W_u e W_e sono due forme asintoticamente equivalenti al primo ordine a $W(\theta)$ (cfr. ad esempio Azzalini, 1992, Cap. 5 o Pace e Salvan, 1996, Cap. 3).

Un modello viene detto **modello statistico parametrico regolare** quando:

- il supporto non dipende da θ;
- è soddisfatta la condizione di identificabilità;
- il modello è correttamente specificato, ossia $p^0(y) = p(y;\theta_0)$, per un valore $\theta_0 \in \Theta$ che indica il vero valore del parametro;
- la funzione di log-verosimiglianza ammette in un intorno di θ_0 uno sviluppo di Taylor fino al secondo ordine e con valore assoluto del resto uniformemente maggiorabile (pertanto $l(\theta)$ è derivabile fino al terzo ordine con valore assoluto delle derivate terze maggiorabile);
- per $l(\theta)$ e le sue derivate fino al terzo ordine esiste finito il valore atteso rispetto alla distribuzione nulla (in particolare esso è pari a 0 per la *score* e vale l'identità dell'informazione con $i(\theta_0) > 0$);
- esiste finito il valore atteso nullo della funzione di *Y* che maggiora il valore assoluto della derivata terza di $l(\theta)$.

Sotto queste condizioni di regolarità, i risultati asintotici del primo ordine richiedono che la quantità d'informazione disponibile sia grande, nel senso che $i(\theta) = O(n)$. Tale condizione è soddisfatta se i dati y sono costituiti da n

osservazioni indipendenti, per cui $i(\theta) = n\, i_1(\theta)$, dove $i_1(\theta)$ indica l'informazione attesa di Fisher per singola osservazione.

Si assuma per semplicità $p = 1$. Nel caso multiparametrico i risultati di seguito presentati continuano a valere, con le opportune reinterpretazioni. Per il teorema del limite centrale si ha che la distribuzione limite nulla della *score* è normale, ossia

$$i(\theta)^{-1/2} l_*(\theta) \xrightarrow{d} N(0,1), \qquad (3.13)$$

dove il simbolo $\xrightarrow{d}$ indica la convergenza in distribuzione. La (3.13) è la base per stabilire proprietà asintotiche del primo ordine per le altre quantità di verosimiglianza di interesse, ovvero per lo s.m.v. o il log-rapporto di verosimiglianza, e le sue forme asintoticamente equivalenti.

Da uno sviluppo della funzione *score* si ottiene che

$$(\hat{\theta} - \theta) i(\theta)^{1/2} = i(\theta)^{-1/2} l_*(\theta)\left(1 + o_p(1)\right), \qquad (3.14)$$

che dà $(\hat{\theta} - \theta) \xrightarrow{d} N(0, i(\theta)^{-1})$. Si osservi che, come stimatore di θ, $\hat{\theta}$ è asintoticamente non distorto e che la sua varianza asintotica raggiunge la soglia inferiore di Cramer-Rao. Come conseguenza dello sviluppo asintotico (3.14) si ottiene la consistenza debole dello s.m.v. $\hat{\theta}$; più precisamente, si ha che $\hat{\theta} - \theta = O_P(n^{-1/2})$.

Considerando anche uno sviluppo della $l(\theta)$, e sostituendo in esso lo sviluppo di $l_*(\theta)$, risulta

$$W(\theta) = 2\left\{l(\hat{\theta}) - l(\theta)\right\} = \left(\hat{\theta} - \theta\right)^2 i(\theta)\left\{1 + o_p(1)\right\}. \qquad (3.15)$$

Pertanto, in base alla distribuzione asintotica (3.13) di $(\hat{\theta} - \theta)$, si ha il risultato di distribuzione asintotica nulla $W(\theta) \xrightarrow{d} \chi^2_1$. Nel caso di parametro con p componenti, la distribuzione asintotica nulla di W è

$$W(\theta) \xrightarrow{d} \chi^2_p.$$

I risultati sulla distribuzione asintotica di *score* e s.m.v. (3.13) e (3.14) comportano che anche W_u e W_e abbiano distribuzione nulla asintotica χ^2_p, per $p \geq 1$, ovvero

$$W_u(\theta) = l_*(\theta)^T i(\theta)^{-1} l_*(\theta) \xrightarrow{d} \chi^2_p$$

$$W_e(\theta) = (\hat{\theta} - \theta)^T i(\theta)(\hat{\theta} - \theta) \xrightarrow{d} \chi^2_p .$$

In realtà, gli sviluppi asintotici (3.13) e (3.14), opportunamente estesi al caso $p \geq 1$, dimostrano che W_u e W_e sono asintoticamente equivalenti al primo ordine a W, essendo $W = W_u + o_p(1)$ e $W = W_e + o_p(1)$.

Se il parametro θ è scalare, ossia p = 1, può essere opportuno fare riferimento alle versioni unilaterali, ossia con segno, date da

$$r(\theta) = \text{sgn}(\hat{\theta} - \theta)\sqrt{W(\theta)} , \tag{3.16}$$

$$r_u(\theta) = l_*(\theta)(i(\theta))^{-\frac{1}{2}} , \tag{3.17}$$

$$r_e(\theta) = (\hat{\theta} - \theta)(i(\theta))^{\frac{1}{2}} , \tag{3.18}$$

che hanno distribuzione asintotica nulla *N*(0,1) con ampia generalità (si veda Pace e Salvan, 1996, Cap. 3). Le distribuzioni asintotiche si mantengono valide se la matrice di informazione attesa di Fisher $i(\theta)$ è stimata utilizzando $i(\hat{\theta})$ o $j(\hat{\theta})$.

3.3. Presenza di parametri di disturbo

In molte situazioni di interesse pratico, il parametro θ può essere partizionato come $\theta = (\psi, \lambda)$, con ψ parametro d'interesse scalare e λ parametro di disturbo (*p*-1)-dimensionale. Per una corretta specificazione del modello è indispensabile tener presente di tutta la struttura probabilistica esaminata, in particolar modo di quella d'interesse senza, però, trascurare i parametri di disturbo che permettono di individuare meglio la variabilità del problema.

Si possono avere due partizioni del parametro θ:

D. il parametro di disturbo λ e quello d'interesse ψ sono comuni a tutte le osservazioni (i.i.d), e perciò non dipendono dalla numerosità campionaria;
E. il parametro di disturbo λ dipende dalla numerosità campionaria, ossia $\lambda = (\lambda_1, ..., \lambda_n)$, mentre quello d'interesse è comune a tutte le osservazioni.

   In questo caso i parametri di disturbo sono detti parametri incidentali.

Nel seguito si farà riferimento alla situazione 1., con $\theta = (\psi, \lambda)$. La log-verosimiglianza assume la forma $l(\theta) = l(\psi, \lambda)$ e la *score* è partizionata come

$$l_* = l_*(\theta) = \begin{pmatrix} l_\psi(\theta) \\ l_\lambda(\theta) \end{pmatrix}, \tag{3.19}$$

dove $l_\psi = l_\psi(\theta) = \partial l(\psi, \lambda)/\partial \psi$ e, similmente, $l_\lambda = l_\lambda(\theta) = \partial l(\psi, \lambda)/\partial \lambda$.

La matrice d'informazione di Fisher è analogamente partizionata come

$$i = i(\theta) = \begin{pmatrix} i_{\psi\psi}(\theta) & i_{\psi\lambda}(\theta) \\ i_{\lambda\psi}(\theta) & i_{\lambda\lambda}(\theta) \end{pmatrix}, \tag{3.20}$$

in analogia con la partizione di l_* nelle componenti l_ψ e l_λ.

In presenza di parametri di disturbo sarebbe conveniente poter basare l'inferenza su una funzione di verosimiglianza che dipenda solo dal parametro di interesse ψ. Tale riduzione di complessità del problema inferenziale risulta tanto più vantaggiosa quanto maggiore è la dimensione del parametro di disturbo λ, soprattutto se la perdita d'informazione su ψ è nulla o trascurabile. Una qualunque funzione dipendente solo dal parametro d'interesse (oltre che dai dati *y*) che si comporti, sotto uno o più aspetti, come una verosimiglianza in senso proprio viene detta funzione di **pseudo-verosimiglianza** (cfr. ad esempio Pace e Salvan, 1996, § 4.3).

La funzione di pseudo-verosimiglianza più utilizzata per l'inferenza sul parametro d'interesse ψ, in presenza del parametro di disturbo λ, prevede di sostituire λ nella verosimiglianza originaria $L(\psi, \lambda)$, con una sua stima consistente. Più precisamente, la funzione di **verosimiglianza profilo** per ψ è definita come

$$L_P(\psi) = L(\psi, \hat{\lambda}_\psi),$$

dove il valore $\hat{\lambda}_\psi$ è la s.m.v. di λ per ψ fissato. Nei problemi regolari, $\hat{\lambda}_\psi$ è soluzione in λ di $l_\lambda(\psi, \lambda) = 0$.

La verosimiglianza profilo, pur non essendo una verosimiglianza propria, gode di alcune proprietà interessanti, che la assimilano a una verosimiglianza propria. In particolare:

- la s.m.v. profilo coincide con la s.m.v. di ψ basata su $L(\theta)$;
- il log-rapporto di verosimiglianza profilo coincide con il log-rapporto di verosimiglianza basato su $L(\psi, \lambda)$, e analogo risultato vale per le versioni unilaterali del test. Si ha, infatti, $W_P(\psi) = 2\left(l_P(\hat{\psi}) - l_P(\psi)\right) = 2\left(l(\hat{\theta}) - l(\psi, \hat{\lambda}_\psi)\right)$;
- si può calcolare un intervallo di confidenza per il parametro ψ di livello asintotico $1-\alpha$ per ogni valore di λ come $\left\{ \psi : W_P(\psi) \leq \chi^2_{1,1-\alpha} \right\}$, con $\chi^2_{1,1-\alpha}$ pari all'($1-\alpha$)-esimo percentile della distribuzione χ^2_1;
- l'informazione osservata profilo è

$$j_P(\psi) = -\frac{\partial^2 l_P(\psi)}{\partial \psi^2} = -\frac{\partial^2}{\partial \psi^2} l(\psi, \hat{\lambda}_\psi). \tag{3.21}$$

Si può mostrare che

$$j_P(\psi) = -(\tilde{l}_{\psi\psi} - \tilde{l}_{\psi\lambda}(\tilde{l}_{\lambda\lambda})^{-1}\tilde{l}_{\lambda\psi}), \tag{3.22}$$

dove il simbolo "~" indica che le quantità di verosimiglianza sono valutate in $(\psi, \hat{\lambda}_\psi)$. Pertanto

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

dove $j^{\psi\psi}(\psi, \lambda)$ rappresenta l'elemento (ψ, ψ) dell'inversa della matrice di informazione osservata complessiva.

Queste proprietà rendono la verosimiglianza profilo interessante. Ma la $L_P(\psi)$ non è una verosimiglianza in senso proprio. In particolare, la funzione *score* profilo non ha valore atteso nullo pari a zero.

Le versioni asintoticamente equivalenti a $W_P(\psi)$ sono

$$W_{eP} = W_{eP}(\psi) = l_*(\psi, \hat{\lambda}_\psi)^2 i_{\psi\psi}(\hat{\psi}, \hat{\lambda}) \xrightarrow{d} \chi^2_1,$$

$$W_{uP} = W_{uP}(\psi) = (\hat{\psi} - \psi)^2 \left\{ j_{\psi\psi}(\hat{\psi}, \hat{\lambda}) \right\}^{-1} \xrightarrow{d} \chi^2_1.$$

Le principali lacune della verosimiglianza profilo sono due:

4) nel caso di parametri incidentali, può non essere appropriato comportarsi come se λ fosse noto e pari a $\hat{\lambda}_\psi$, tipicamente se la dimensione di λ è elevata;
5) nel caso in cui la dimensione del campione osservato è modesta.

A seguito di queste lacune, negli anni recenti sono state elaborate varie versioni modificate della verosimiglianza profilo.

Esempio 3.2: Nel caso in cui il parametro naturale delle famiglie esponenziali è partizionato come $\theta = (\psi, \lambda)$, si ha

$$f(y; \psi, \lambda) = h(y) \exp\left\{ \psi\, v(y) + \lambda^T u(y) - K(\psi, \lambda) \right\}, \qquad (3.23)$$

con $v(y)$ di dimensione 1 e $u(y)$ di dimensione p-1.

Per l'inferenza su ψ si ha che (cfr. ad esempio Pace e Salvan, 1996, § 5.4) la distribuzione condizionata di v dato da u è ancora una famiglia esponenziale, indipendente da λ, con

$$f_{V|U=u}(v; u, \psi) = h_u(v) \exp\left\{ \psi\, v - K_u(\psi) \right\},$$

dove $K_u(\psi)$ dipende da ψ solamente. Questo risultato ha delle implicazioni importanti in quanto permette di definire una **verosimiglianza condizionata** per ψ, data da

$$L_c(\psi) = \exp\left\{ \psi\, v - K_u(\psi) \right\}. \qquad (3.24)$$

Tuttavia la (3.24) esprime in generale un risultato teorico in quanto la forma della funzione $K_u(\psi)$ non è desumibile in modo immediato. Un'eccezione si ha quando la distribuzione marginale di u è nota.

Se non è agevole ottenere la verosimiglianza condizionata esatta, una possibilità è ricorrere alla verosimiglianza profilo, data da $l_P(\psi) = \psi\, v + \hat{\lambda}_\psi^T u - K(\psi, \hat{\lambda}_\psi)$. La *score* profilo è $l_{P*} = v - \tilde{K}_\psi$ e l'informazione profilo è $j_P(\psi) = \tilde{K}_{\psi\psi} - \tilde{K}_{\psi\lambda}\left(\tilde{K}_{\lambda\lambda}\right)^{-1}\tilde{K}_{\lambda\psi}$.

Qualora non sia possibile ricorrere alla verosimiglianza condizionata, che è una verosimiglianza propria, considerazioni di carattere asintotico possono suggerire miglioramenti di $L_p(\psi)$, attraverso l'introduzione di fattori di modificazione. Varie proposte di modificazioni di $L_p(\psi)$ sono state discusse in anni recenti (vedi ad esempio Severini, 2000, Cap. 9). In generale una **verosimiglianza profilo modificata** è definita come

$$L_{PM}(\psi) = L_P(\psi) \cdot M(\psi), \tag{3.25}$$

dove $M(\psi)$ rappresenta un opportuno fattore di aggiustamento di ordine $O_P(1)$. In letteratura sono state proposte diverse espressioni di questo fattore, a partire da Barndorff-Nielsen (1980, 1983). Assumendo che la statistica sufficiente minimale sia esprimibile come $(\hat{\psi}, \hat{\lambda}, a)$, dove a è una statistica ancillare, esattamente o approssimativamente, così che $l(\psi, \lambda; y) = l(\psi, \lambda; \hat{\psi}, \hat{\lambda}, a)$, allora la **log-verosimiglianza profilo modificata** è

$$l_{PM}(\psi) = l_{PM}(\psi; y) = l_P(\psi) - \frac{1}{2}\log |\, j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\,| - \log\left|\frac{\partial \hat{\lambda}_\psi}{\partial \hat{\lambda}}\right|, \tag{3.27}$$

dove $\left|\frac{\partial\hat{\lambda}_\psi}{\partial\hat{\lambda}}\right| = \frac{|\,l_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)\,|}{|\,j_{\lambda\lambda}(\hat{\theta}_\psi)\,|}$, con $l_{\lambda;\hat{\lambda}}(\psi,\lambda) = \frac{\partial^2 l(\psi,\lambda;\hat{\psi},\hat{\lambda},a)}{\partial\lambda\partial\hat{\lambda}^T}$. Quando ψ e λ sono ortogonali, ossia $i_{\psi\lambda}(\theta)=0$, si ha che $\log\left|\frac{\partial\hat{\lambda}_\psi}{\partial\hat{\lambda}}\right| = O_P(n^{-1})$ con $\psi-\hat{\psi}=O_P(n^{-1/2})$.

Si perviene così a definire la log-verosimiglianza condizionata di Cox e Reid (1987), data da

$$l_A(\psi) = l_A(\psi;y) = l_P(\psi) - \frac{1}{2}\log|\,j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)\,|,$$

che approssima $l_{PM}(\psi)$ con errore di ordine $O_P(n^{-1})$.

Una approssimazione di $l_{PM}(\psi)$ sviluppata in Severini (2000) è

$$\tilde{l}_M(\psi) = \tilde{l}_M(\psi;y) = l_P(\psi) + \frac{1}{2}\log|\,j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)\,| - \log\left|v_{\lambda,\lambda}(\psi,\hat{\lambda}_\psi,\hat{\theta};\hat{\theta})\right|, \quad (3.28)$$

dove $v_{\lambda,\lambda}(\psi,\lambda;\theta_0) = E_{\theta_0}(l_\lambda(\psi)l_\lambda(\lambda)^T)$ e $\theta_0=(\psi_0,\lambda_0)$ denota il vero valore del parametro. Un'altra versione della (3.28) è ottenuta sostituendo $v_{\lambda,\lambda}(\psi,\hat{\lambda}_\psi,\hat{\theta};\hat{\theta})$ con l'analogo empirico $\hat{v}_{\lambda,\lambda}(\psi,\hat{\lambda}_\psi,\hat{\theta}) = \sum_{i=1}^{n} l_\lambda(\psi,\lambda,y_i)l_\lambda(\psi,\hat{\lambda}_\psi,y_i)^T$.

Per altre espressioni della log-verosimiglianza profilo modificata si vedano Severini (2000), Pace e Salvan (1996) e Brazzale *et al*. (2007), e i riferimenti qui riportati.

3.4. Metodi asintotici

Come il teorema del limite centrale permette di ottenere le usuali approssimazioni asintotiche per quantità di verosimiglianza, i risultati sulle approssimazioni di ordine più elevato consentono di sviluppare metodi asintotici di ordine superiore per l'inferenza basata sulla verosimiglianza.

In anni recenti sono stati elaborati degli strumenti per ottenere specifici raffinamenti dei risultati di primo ordine. Questi raffinamenti riguardano in particolare le distribuzioni nulle di W e r, anche in presenza di parametri di disturbo, e opportune modificazioni della verosimiglianza profilo.

Nel seguito sono presentate le versioni modificate del test radice con segno di $W(\theta)$ e $W_P(\psi)$ (cfr. Pace e Salvan, 1996, § 11.5).

Per θ scalare, ovvero in assenza di parametri di disturbo, conviene considerare per l'inferenza la quantità $r(\theta)=\text{sgn}(\hat{\theta}-\theta)\sqrt{W(\theta)}$. Per ottenere una maggiore accuratezza nelle approssimazioni, si può ricorrere alla **radice con segno modificata del log-rapporto di verosimiglianza**, r^*, introdotta da Barndorff-Nielsen (1980, 1983). Tale quantità è data da

$$r^* = r + r^{-1}\log\frac{U}{r}\,, \qquad (3.28)$$

dove U è invariante rispetto alla riparametrizzazione ed è dato da

$$U = j(\hat{\theta})^{-1/2}(\hat{l}_{;1}-l_{;1})\,, \qquad (3.29)$$

con $l_{;1}=\dfrac{\partial}{\partial\hat{\theta}}l(\theta;\hat{\theta},a)$ e il simbolo "^" indica che le quantità di verosimiglianza sono valutate in $\hat{\theta}$. Allora vale che

$$P_\theta(r^*\le r\mid a)=\Phi(r)\left\{1+O\left(n^{-3/2}\right)\right\},$$

e pertanto r^* ha distribuzione normale $N(0,1)$ con errore di ordine $O(n^{-3/2})$ sia condizionatamente ad a sia marginalmente. Tale risultato rende più veloce la convergenza di r^* alla propria distribuzione asintotica di quanto accade per $r(\theta)$.

<u>**Esempio 3.1 (cont.)**</u>: Sia y un campione casuale semplice con densità

$$f(y;\theta)=h(y)\exp\{\theta\, y-K(\theta)\}.$$

Le quantità necessarie per il calcolo di r^* sono $l(\theta;\hat{\theta})=\theta\,K'(\hat{\theta})-K(\theta)$, $l_{;1}(\theta)=\theta\,K''(\hat{\theta})=\theta\,j(\hat{\theta})$ e $\hat{l}_{;1}=\hat{\theta}\,K''(\hat{\theta})=\hat{\theta}\,j(\hat{\theta})$.

Si ottiene quindi

$$U = j(\hat{\theta})^{-1/2}(\hat{l}_{;1} - l_{;1}) = j(\hat{\theta})^{1/2}(\hat{\theta} - \theta) = \text{sgn}(\hat{\theta} - \theta)\sqrt{j(\hat{\theta})(\hat{\theta} - \theta)^2} \ .$$

Si osservi che U coincide con r_e, ossia la statistica di Wald con segno nella riparametrizzazione naturale.

<u>Esempio 3.3</u>: Un caso particolare della famiglia esponenziale è rappresentato dalla distribuzione *Bin (n,p).* Per l'inferenza su *p*, uno degli approcci classici consiste nell'usare la statistica radice con segno del log-rapporto di verosimiglianza.

$$r = r(p) = \text{sgn}(\hat{p} - p)\sqrt{2(l(\hat{p}) - l(p))} = \text{sgn}(\hat{p} - p)\sqrt{2n\hat{p}\log\frac{\hat{p}}{1-\hat{p}} + 2n\log\frac{1-\hat{p}}{1-p}} \ ,$$

dove $l(p) = n\hat{p}\log\left(\dfrac{p}{1-p}\right) + n\log(1-p)$ denota la funzione di log-verosimiglianza per *p*. La statistica *r* è asintoticamente distribuita secondo una distribuzione normale standard. E' possibile considerare la versione modificata della statistica radice con segno del log-rapporto di verosimiglianza, *r**, che ha una distribuzione normale con approssimazione di ordine superiore. La versione modificata di *r* è data dalla (3.28), con *U* termine di correzione definito come

$$U = U(p) = \sqrt{n\hat{p}(1-\hat{p})}\left(1 - \frac{(1-\hat{p})p}{\hat{p}(1-p)}\right).$$

Si deve tener presente che r^* non può essere utilizzato per $p = 1$. Infatti, $u = 0$ e r è finito; rappresenta un caso limite.

Si consideri il caso multiparametrico con $\theta = (\psi, \lambda)$, con ψ scalare. Si può ottenere (Barndorff-Nielsen, 1991b) una versione modificata di $r_P = \text{sgn}(\hat{\psi} - \psi)\sqrt{W_P}$, con $W_P = W_P(\psi) = 2\left(l_p(\hat{\psi}) - l_P(\psi)\right) = 2\left(l(\hat{\theta}) - l(\psi, \hat{\lambda}_\psi)\right)$, avente distribuzione nulla $N(0,1)$ con errore di ordine $O\left(n^{-3/2}\right)$. Si assume, inoltre, che la statistica sufficiente minimale sia esprimibile nella forma $(\hat{\theta}, a) = (\hat{\psi}, \hat{\lambda}, a)$ con *a* ancillare. La versione modificata di r_P è definita come

$$r_P{*} = r_P + \frac{1}{r_P}\log\frac{CU_P}{r_P}\,, \qquad (3.31)$$

con

$$C = \frac{|\,\tilde{l}_{\lambda;\hat{\lambda}}\,|}{\left\{|\,\tilde{j}_{\lambda\lambda}\,||\,\hat{j}_{\lambda\lambda}\,|\right\}^{1/2}}\,, \qquad (3.32)$$

e

$$U_P = j_P(\hat{\psi})^{-1/2}\frac{\partial}{\partial\hat{\psi}}\left\{l_p(\psi) - l_P(\hat{\psi})\right\}. \qquad (3.33)$$

Un generico elemento della matrice $l_{\lambda;\hat{\lambda}}$ nella (3.32) è

$$l_{a;b} = \frac{\partial^2 l(\psi,\lambda;\hat{\psi},\hat{\lambda},a)}{\partial\lambda^a\partial\hat{\lambda}^b}, \text{ con } a,b = 1,\ldots,p-1. \qquad (3.34)$$

La derivata parziale che compare nella (3.33) va intesa come una derivata rispetto a $\hat{\psi}$ della log-verosimiglianza profilo normalizzata, considerata come funzione di ψ, $\hat{\psi}$, $\hat{\lambda}_\psi$ e di a. Tale derivata può essere scritta, anche, nella forma

$$\frac{\partial}{\partial\hat{\psi}}\left\{l_P(\psi\;;\hat{\psi},\hat{\lambda},a) - l_P(\hat{\psi}\;;\hat{\psi},\hat{\lambda},a)\right\} = \tilde{l}_{;\hat{\psi}} - \hat{l}_{;\hat{\psi}} + \frac{\partial\hat{\lambda}}{\partial\hat{\psi}}(\tilde{l}_{;\hat{\lambda}} - \hat{l}_{;\hat{\lambda}})\,, \qquad (3.35)$$

dove la quantità $\frac{\partial\hat{\lambda}}{\partial\hat{\psi}}$ si ottiene derivando rispetto a $\hat{\psi}$ l'equazione di verosimiglianza per $\hat{\lambda}_\psi$. Risulta $\frac{\partial\hat{\lambda}}{\partial\hat{\psi}} = -(\tilde{l}_{\lambda;\hat{\lambda}})^{-1}\tilde{l}_{\lambda;\hat{\psi}}$.

<u>Esempio 3.2 (cont.)</u>: Sia y un campione casuale semplice con densità (cfr. Esempio 3.2)

$$f(y;\psi,\lambda) = h(y)\exp\left\{\psi\, v(y) + \lambda^T u(y) - K(\psi,\lambda)\right\},$$

con funzione di log-verosimiglianza

$$l(\psi,\lambda;\hat{\psi},\hat{\lambda}) = \psi\, t + \lambda^T u - K(\psi,\lambda) = \left\{\psi\, K_\psi(\hat{\psi},\hat{\lambda}) + \lambda K_\lambda(\hat{\psi},\hat{\lambda}) - K(\psi,\lambda)\right\},$$

con $K_{\psi}(\psi,\lambda)=\frac{\partial}{\partial\psi}K(\psi,\lambda)$ e, analogamente, $K_{\lambda}(\psi,\lambda)=\frac{\partial}{\partial\lambda}K(\psi,\lambda)$. Si ottiene $l_{\lambda;\hat\lambda}=K_{\lambda\lambda}(\hat\psi,\hat\lambda)$ e $j_{\lambda\lambda}=K_{\lambda\lambda}(\psi,\lambda)$, da cui

$$C=\frac{|K_{\lambda\lambda}(\hat\psi,\hat\lambda)|^{1/2}}{|K_{\lambda\lambda}(\psi,\hat\lambda_{\psi})|^{1/2}}.$$

Inoltre,

$$\tilde l_{;\hat\psi}-\hat l_{;\hat\psi}=(\psi-\hat\psi)K_{\psi\psi}(\hat\psi,\hat\lambda)+(\hat\lambda_{\psi}-\hat\lambda)K_{\lambda\psi}(\hat\psi,\hat\lambda),$$

$$\frac{\partial\hat\lambda}{\partial\hat\psi}=-\left(K_{\lambda\lambda}(\hat\psi,\hat\lambda)\right)^{-1}K_{\lambda\psi}(\hat\psi,\hat\lambda),$$

$$\tilde l_{;\hat\lambda}-\hat l_{;\hat\lambda}=(\psi-\hat\psi)K_{\psi\lambda}(\hat\psi,\hat\lambda)+(\hat\lambda_{\psi}-\hat\lambda)K_{\lambda\lambda}(\hat\psi,\hat\lambda).$$

Pertanto, tenuto conto che $j_P(\psi)=\tilde K_{\psi\psi}-\tilde K_{\psi\lambda}\left[\tilde K_{\lambda\lambda}\right]^{-1}\tilde K_{\lambda\psi}$, si può verificare che U_P è equivalente a $v_P=(\hat\psi-\psi)\,j_P(\hat\psi)^{-1/2}$.

Si ricordi che nell'ambito delle famiglie esponenziali, il parametro di disturbo λ può essere eliminato tramite condizionamento. Inoltre, si ha che

$$f_{V|U=u}(v;u,\psi)=\exp\{l_{MP}(\psi)\}\{1+O(n^{-1})\},$$

dove $l_{MP}(\psi)=l_P(\psi)+\frac{1}{2}\log|\,j_{\lambda\lambda}(\psi,\hat\lambda_{\psi})|-\log|\,l_{\lambda;\hat\lambda}(\psi,\hat\lambda_{\psi};\hat\psi,\hat\lambda,a)|$. Utilizzando la versione profilo modificata, si può ottenere una statistica di ordine elevato, prendendo

$$r=\operatorname{sgn}(U)\left[2\{l_{MP}(\hat\psi_{MP})-l_{MP}(\psi)\}\right]^{1/2},$$

con

$$U=j_{MP}(\hat\psi_{MP})^{-1/2}(\hat\psi_{MP}-\psi),$$

con $\hat\psi_{MP}$ s.m.v. profilo modificata e $j_{MP}(\psi)$ informazione osservata di Fisher per la profilo modificata.

3.5. Considerazioni conclusive

Questo capitolo di rassegna desidera introdurre la teoria classica e avanzata che verrà applicata al dataset, oggetto di studio in questa tesi, nel capitolo successivo. Obiettivo è, quindi, fornire procedure di inferenza utilizzate nel seguito.

Il Capitolo 4, oltre ad applicare ai dati procedure inferenziali di base, utilizza le recenti versioni modificate della statistica radice con segno del log-rapporto di verosimiglianza per ottenere risultati più precisi anche con bassa numerosità campionaria. Queste modificazioni permettono, inoltre, una convergenza più rapida verso la distribuzione asintotica migliorandone la precisione.

Capitolo 4

Metodi asintotici di verosimiglianza e il caso studio di Padova

In questo capitolo si utilizzano le procedure di inferenza, introdotte nei capitoli precedenti, al caso di studio di Padova, presentato nel Capitolo 2.
Si ricorda che le variabili d'interesse sono quelle relative ai nuovi test diagnostici che vogliamo testare, ossia **PCR MB A** e **PCR MB B**. Si vorranno confrontare tali tecniche con i risultati ottenuti dal test diagnostico classico, ossia la **Coltura**.
Verrà considerata anche la variabile Colorazione Z-N.
All'inizio del capitolo sono richiamati i concetti di specificità e sensibilità introdotti nel Capitolo 1, e tali quantità vengono calcolate per le variabili di interesse. Viene fornita una loro stima puntuale ed intervallare, utilizzando per quest'ultima le tre possibili espressioni viste nel Capitolo 1: intervallo di confidenza di Wald, intervallo di confidenza *score* ed intervallo di confidenza di Wald aggiustato.

Inoltre, vengono applicate le procedure di inferenza di ordine elevato discusse nel Capitolo 3 ai dati disponibili. Queste procedure rappresentano metodi asintotici basati sulla verosimiglianza per l'inferenza su quantità di interesse, che forniscono risultati accurati anche se la numerosità, come nel nostro campione, è piccola o moderata.
Per la stesura di questo capitolo si è principalmente fatto riferimento ai libri di Bortot *et al.* (2000) e di Piccolo (1998) per la parte di teoria classica; mentre, per la parte di procedure di inferenza basate sulla verosimilgianza di ordine superiore, al volume Brazzale *et al.* (2007, § 3.4, 4.3 e 4.1).

4.1 Sensibilità e specificità dei nuovi test diagnostici

Per la valutazione di un test diagnostico e, nello specifico, per calcolarne l'affidabilità, risultano essere molto utili le due quantità introdotte nel Capitolo 1: sensibilità e specificità.
Si desidera calcolare tali quantità per i due nuovi test diagnostici (ovvero per entrambe le Tecniche A e B) con riferimento alla variabile **Coltura**. Con riferimento alla Tabella 1.1, nel nostro caso la malattia è identificata dalla variabile **Coltura** che distingue i pazienti in sani/malati. L'esito del test, invece, è costituito dai risultati sui test diagnostici da valutare. Per ogni incrocio sono stati, però, esclusi i pazienti che in almeno in una delle due variabili risultavano assumere la modalità "assente". Questo riduce la dimensione campionaria.

Nella Tabella 4.1 è riportato l'incrocio tra le variabili **Coltura** e **PCR MB A**. La sensibilità per la **PCR MB A** risulta essere del 25%, mentre la specificità è del 100%. Questo significa che il test PCR con la Tecnica A non è molto efficace nell'identificare i veri positivi, contrariamente ai risultati negativi. Questi valori di SN e SP sono simili a quelli assunti dall'esame di microscopia diretta riportati in letteratura, ossia 30 % e 80% rispettivamente (vedi Capitolo 1).

		COLTURA		
PCR MB A		NEGATIVA	POSITIVA	TOTALE PCR MB A
	NEGATIVA	7	3	10
	POSITIVA	0	1	1
	TOTALE COLTURA	7	4	11

Tabella 4.1: *Confronto tra* ***PCR MB A*** *e* ***Coltura****.*

Nella Tabella 4.2 sono riportate le stime puntuali ed intervallari per SN e SP della tecnica diagnostica **PCR MB A**.

	Sensibilità	Specificità
Wald	0.25 (0.00, 0.67)	1.00 non definito
Score	0.25 (0.05, 0.70)	1.00 (0.65, 1.00)
Wald aggiustato	0.37 (0.04, 0.71)	0.82 (0.59, 1.00)

Tabella 4.2: *Stime puntuali ed intervallari per sensibilità e specificità calcolati per* ***PCR MB A****.*

Una analisi equivalente può essere fatta per il confronto tra **PCR MB B** su micobatterio con la Tecnica B e **Coltura** (Tabella 4.3). In questo caso la sensibilità è dello 0%, ossia i due esami non risultano trovare corrispondenza per gli esiti positivi; invece, la specificità è del 100%. Notiamo che questi indici sono peggiori rispetto a quelli relativi alla Tecnica A e, anche, rispetto agli indici riportati in letteratura per l'esame microscopico diretto.

		COLTURA		
PCR MB B		**NEGATIVA**	**POSITIVA**	**TOTALE PCR MB B**
	NEGATIVA	10	4	14
	POSITIVA	0	0	0
	TOTALE COLTURA	10	4	14

Tabella 4.3: *Confronto tra* ***PCR MB B*** *e* ***Coltura****.*

Nella Tabella 4.4 vengono riportate le stime puntuali e intervallari di sensibilità e specificità della tecnica diagnostica **PCR MB B**.

	Sensibilità	**Specificità**
Wald	0.00 non definito	1.00 non definito
Score	0.00 (0.00, 0.49)	1.00 (0.72, 1.00)
Wald aggiustato	0.25 (0.00, 0.55)	0.86 (0.67, 1.00)

Tabella 4.4: *Stime puntuali ed intervallari per sensibilità e specificità calcolati per* ***PCR MB B****.*

I risultati sulla sensibilità e sulla specificità, fanno risaltare come entrambi i test risultano eccellenti nell'identificazione dei veri negativi, mentre non sono affidabili nell'individuazione dei veri positivi. Inoltre, l'intervallo di confidenza di Wald aggiustato (Agresti e Coull, 1998), oltre ad essere molto semplice da calcolare, sembra fornire i risultati più accurati. Anche se con dimensioni campionarie diverse, si può osservare che la Tecnica A sembra preferibile alla Tecnica B in quanto, oltre ad identificare correttamente tutti i veri negativi, individua il 25% dei veri positivi. Considerando, anche, gli intervalli di confidenza tale considerazione resta invariata.

Si consideri, infine, la relazione tra le variabili **Colorazione ZN** e **Coltura** (vedi Tabella 4.5). Come già accennato nel Capitolo 1, la colorazione ZN costituisce un'analisi esplorativa, il cui risultato è immediato, ma non molto affidabile (sicuramente meno affidabile della Coltura). Nella Tabella 4.6 sono riportate le stime puntuali e intervallari di SN e SP per la variabile **Colorazione ZN**.

		COLTURA		
COLORTAZIONE ZN		**NEGATIVA**	**POSITIVA**	**TOTALE COLORAZIONE ZN**
	NEGATIVA	12	4	16
	POSITIVA	1	1	2
	TOTALE COLTURA	13	5	18

Tabella 4.5: *Confronto tra **Colorazione ZN** e **Coltura**.*

Stime puntuali (i.c.)	**Sensibilità**	**Specificità**
Wald	0.20 (0.00, 0.55)	0.92 (0.78,1.00)
Score	0.20 (0.04, 0.62)	0.92 (0.67, 0.99)
Wald aggiustato	0.33 (0.03, 0.64)	0.82 (0.64, 1.00)

Tabella 4.6: *Stime puntuali ed intervallari per sensibilità e specificità per* ***Colorazione ZN***.

La sensibilità e la specificità assumono, rispettivamente, i valori: 20% e 92%, o 0.33 e 0.82 se calcolati secondo la regola suggerita da Agresti e Coull (1998). Notiamo che nella Tabella 4.5 la numerosità campionaria è maggiore rispetto alle analisi precedenti. Il valore della sensibilità resta, tuttavia, piuttosto basso; mentre la specificità è alta come nei risultati precedenti. Si può osservare, inoltre, una corrispondenza tra i valori assunti da SN e SP per la variabile **PCR MB A** e quelli relativi alla **Colorazione ZN**; queste due variabili, infatti, come sottolineato nel test esatto di Fisher sono correlate (vedi Tabella 2.8).

4.2 Analisi tabelle 2x2

Molto spesso, nelle applicazioni come quella di interesse in questa tesi, si hanno a disposizione dati classificati secondo una tabella di contingenza, in genere una 2x2. Questo è particolarmente frequente nei dataset di natura medica o nelle scienze sociali, per le quali è frequente la formazione di gruppi attorno a variabili dicotomiche.

Si consideri la situazione in cui la variabile che identifica la presenza di micobatteri tramite un test diagnostico (ad esempio PCR o Colorazione ZN) venga modellata come una variabile binomiale $Y_0 \sim Bi(m_0, p_0)$ nel gruppo risultato positivo alla coltura e con la variabile $Y_1 \sim Bi(m_1, p_1)$, indipendente da Y_0, nel gruppo risultato negativo alla coltura. Si desidera confrontare p_0 e p_1, e questa comparazione può essere eseguita su varie scale. In particolare, se si considera il log-rapporto delle quote (*log-odds ratio*, vedi § 1.2), il parametro di interesse per l'inferenza è

$$\psi = \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_0}{1-p_0}\right). \tag{4.1}$$

Il parametro (4.1) è il parametro canonico nella famiglia esponenziale

$$f(y;p) = \binom{m_1}{y_1} p_1^{y_1}(1-p_1)^{m_1-y_1} \binom{m_0}{y_0} p_0^{y_0}(1-p_0)^{m_0-y_0}, \text{ con } p = (p_0, p_1).$$

Infatti, possiamo riscrivere questa densità come

$$f(y;p) = \binom{m_1}{y_1}\binom{m_0}{y_0} \exp\left\{y_1\psi + (y_1+y_0)\lambda - m_0\log\left(1+e^{\lambda}\right) - m_1\log\left(1+e^{\lambda+\psi}\right)\right\},$$

con $\lambda = \log\dfrac{p_0}{1-p_0}$.

I parametri λ e ψ sono indipendenti e assumono valori reali, con

$$p_0 = \frac{e^{\lambda}}{1+e^{\lambda}} \text{ e } p_1 = \frac{e^{\lambda+\psi}}{1+e^{\lambda+\psi}}.$$

Si osservi che il parametro di interesse ψ può essere interpretato come il coefficiente angolare in un modello di regressione logistica per la variabile PCR (o colorazione ZN) con una variabile esplicativa dicotomica (coltura); si veda, ad esempio, Pace e Salvan (1996, p. 270).

Per l'inferenza su ψ si possono considerare le diverse procedure presentate nel Capitolo 3: il test alla Wald profilo e quello basato sulla verosimiglianza condizionata, la statistica r_P^* e quella basata sulla verosimiglianza condizionata. Tutte queste procedure sono implementate nella libreria HOA di R (cfr. Brazzale *et al.*, 2007) e possono essere facilmente applicate al nostro caso di studio.

Si presentano di seguito i risultati di tali procedure per le variabili **PCR MB A** e **Colorazione ZN** rispetto alla classificazione della **Coltura**. In queste analisi, per evitare la presenza di celle vuote, che ostacolano il calcolo delle statistiche, sono stati utilizzati dei valori "aggiustati"; si veda per approfondimenti Brazzale e Davison (2009). Inoltre, non viene considerata la variabile **PCR MB B** sia a causa dei risultati precedenti su SP e SN, per cui non si è dimostrato un esame affidabile, sia perché la tabella corrispondente contiene un'intera riga senza valori e questo ne impedisce il calcolo.

Per quest'analisi, è di interesse valutare l'ipotesi nulla $H_0 : \psi = 0$ che indica l'assenza di un effetto significativo della variabile **Coltura** sull'esito del test diagnostico.

Nel primo caso, consideriamo la variabile **PCR MB A** e **Coltura**, i risultati sono riportati nel Codice 4.1.

```
> tab1<-data.frame(n=c(6.5,0.5),t=c(10,1),x=c(0,1))
> tab1
    n  t x
1 6.5 10 0
2 0.5  1 1
> summary(tab1.fit<-glm (cbind(n,t-n)~x,binomial,data=tab1))

Call:
glm(formula = cbind(n, t - n) ~ x, family = binomial, data = tab1)

Deviance Residuals:
[1]  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.619      0.663   0.934    0.350
x             -0.619      2.107  -0.294    0.769

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.085372  on 1  degrees of freedom
Residual deviance: 0.000000  on 0  degrees of freedom
AIC: 8.26

Number of Fisher Scoring iterations: 3

> summary(cond.glm(tab1.fit,offset=x),test=0)

 Formula:  cbind(n, t - n) ~ x
```

```
 Family:  binomial
 Offset:  x

          Estimate   Std. Error
uncond.     -0.6190        2.107
cond.       -0.5642        2.021

Test statistics
---------------
 hypothesis : coef( x ) = 0
                                         statistic   tail prob.
Wald pivot                                  -0.2938       0.3845
Wald pivot (cond. MLE)                      -0.2792       0.3901
Likelihood root                             -0.2922       0.3851
Modified likelihood root                    -0.3376       0.3678
Modified likelihood root (cont. corr.)       0.6672       0.2523

"q" correction term: -0.2927

Diagnostics:
-----------
    INF      NP
0.15145 0.06004

 Approximation based on 20 points
```

Codice 4.1: *Inferenza condizionata approssimata per la **PCR MB A**.*

Tutte le statistiche calcolate per **PCR MB A** portano al rifiuto dell'ipotesi nulla; perciò l'esito nella variabile **Coltura** non influenza l'esito nel test diagnostico PCR MB A.

Nel secondo caso, consideriamo la variabile **Colorazione ZN** e **Coltura**, i risultati sono riportati nel Codice 4.2.

```
> tab3<-data.frame(n=c(12,1),t=c(16,2),x=c(0,1))
> tab3
   n  t x
1 12 16 0
2  1  2 1
> summary(tab3.fit<-glm (cbind(n,t-n)~x,binomial,data=tab3))

Call:
```

```
glm(formula = cbind(n, t - n) ~ x, family = binomial, data = tab3)

Deviance Residuals:
[1]  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0986     0.5774   1.903   0.0571 .
x            -1.0986     1.5275  -0.719   0.4720
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5.0301e-01  on 1  degrees of freedom
Residual deviance: 1.7764e-15  on 0  degrees of freedom
AIC: 8.3678

Number of Fisher Scoring iterations: 3

> summary(cond.glm(tab3.fit,offset=x),test=0)

 Formula:  cbind(n, t - n) ~ x
 Family:  binomial
 Offset:  x

          Estimate   Std. Error
uncond.      -1.099        1.528
cond.        -1.032        1.477

Test statistics
---------------
 hypothesis : coef( x ) = 0
                                         statistic   tail prob.
Wald pivot                                -0.719200      0.2360
Wald pivot (cond. MLE)                    -0.698300      0.2425
Likelihood root                           -0.709200      0.2391
Modified likelihood root                  -0.706600      0.2399
Modified likelihood root (cont. corr.)    -0.002259      0.4991

"q" correction term: -0.7081

Diagnostics:
-----------
    INF      NP
0.08678 0.08116

 Approximation based on 20 points
```

Codice 4.2: *Inferenza condizionata approssimata per la* ***Colorazione ZN****.*

Anche per il Codice 4.2 emergono conclusioni analoghe a quelle relative al test PCR. Infatti, anche per la **Colorazione ZN** si accetta l'ipotesi nulla: la Coltura non ha un legame significativo nemmeno con questa variabile.

Analizziamo, di seguito, la relazione tra la variabile **PCRMB A** e **Colorazione ZN**.

```
tab4<-data.frame(n=c(13,3),t=c(14,4),x=c(0,1))
> tab4
   n  t x
1 13 14 0
2  3  4 1
> summary(tab4.fit<-glm (cbind(n,t-n)~x,binomial,data=tab4))

Call:
glm(formula = cbind(n, t - n) ~ x, family = binomial, data = tab4)

Deviance Residuals:
[1]  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.565      1.038   2.472   0.0134 *
x             -1.466      1.552  -0.945   0.3449
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8.5435e-01  on 1  degrees of freedom
Residual deviance: 1.7764e-15  on 0  degrees of freedom
AIC: 7.6529

Number of Fisher Scoring iterations: 4

> summary(cond.glm(tab4.fit,offset=x),test=0)

 Formula:  cbind(n, t - n) ~ x
 Family:  binomial
 Offset:  x

          Estimate   Std. Error
uncond.      -1.466        1.552
cond.        -1.368        1.490

Test statistics
---------------
```

```
 hypothesis : coef( x ) = 0
                                          statistic   tail prob.
Wald pivot                                   -0.9445      0.1725
Wald pivot (cond. MLE)                       -0.9181      0.1793
Likelihood root                              -0.9243      0.1777
Modified likelihood root                     -0.9163      0.1798
Modified likelihood root (cont. corr.)       -0.2183      0.4136

"q" correction term: -0.9178

Diagnostics:
-----------
    INF      NP
0.08452 0.11295

 Approximation based on 20 points
```

Anche per quest'analisi non si rilevano indicazioni significative per il rifiuto dell'ipotesi nulla.

Da queste ultime analisi, possiamo notare che la statistica r_P* (`Modified likelihood root (cont. corr.)`) basata sulla verosimiglianza condizionata assume valori diversi rispetto alle altre statistiche in quanto tiene conto della bassa numerosità campionaria ed è quindi più attendibile per il nostro caso.

4.3 Considerazioni conclusive

In questo capitolo, sono stati analizzati, utilizzando diverse procedure inferenziali, i dati a disposizione. Applicando la teoria descritta nel Capitolo 1, sono stati calcolati gli indici di sensibilità e specificità, i corrispondenti intervalli di confidenza nelle varie versioni, e applicate, infine, le più recenti procedure introdotte nel Capitolo 3.

I risultati ottenuti per specificità e sensibilità nelle varie tabelle indicano che la **PCR MB A** e **PCR MB B** non identificano molto precisamente i veri positivi; infatti il valore più alto riscontrato è 0.25 con la Tecnica A. Mentre sembrano tutti i test molto efficienti nell'identificare i reali negativi.

Anche con le analisi inferenziali di ordine superiore, sfortunatamente, non si sono rilevate correlazioni significative tra i test diagnostici e la coltura.

Gli obiettivi principali di questa tesi erano: presentare un'analisi statistica utile per studiare e valutare l'efficacia di un nuovo test diagnostico, la PCR (*Polimerase Chain Reaction*), confrontando le due tecniche A e B e presentare le recenti teorie più avanzate sulla verosimiglianza, in particolare le versioni modificate della verosimiglianza profilo.
All'inizio, nel confrontare le due nuove tecniche, sembrava essere preferibile la tecnica B perché questa riusciva ad estrarre DNA dove l'altra tecnica aveva fallito. Mentre continuando l'analisi, si è rilevato che i campioni processati con la Tecnica A erano migliori per l'esecuzione dell'esame PCR su MB. Infatti, con questa tecnicasi è riusciti ad individuare risultati positivi cosa che non succede, invece, per il gruppo B.
Confrontando, invece, la tecnica PCR con la Coltura si giunge alla conclusione che il nuovo test nelle due versioni A e B non risulta molto efficace nell'individuare la presenza/assenza di micobatteri; infatti, fallisce in molti campioni, soprattutto per quelli che realmente manifestano la suddetta patologia.

In conclusione, la tecnica PCR, anche se fornisce gli esiti degli esami in tempi minori rispetto alla coltura, non risulta abbastanza attendibile da sostituirsi a quest'esame; al massimo potrebbe essere proposto, solo, come esame aggiuntivo.
Per la valutazione della tecnica sono state molto utili, anche le quantità di verosimiglianza introdotte recentemente, in quanto forniscono risultati più attendibili in caso di numerosità campionarie ridotte, anche se nel nostro caso non hanno evidenziato legami significativi tra le variabili. Questo studio, comunque, potrebbe essere applicato anche nelle prossime analisi riguardanti la validazione di altri test diagnostici. I vantaggi nell'applicazione di quest'analisi sono, in primo luogo, che i comandi sono già impostati nel programma R facilitandone l'utilizzo e, in secondo luogo, che sarebbe un'ottima occasione per diffondere la statistica anche per applicazioni reali con dati medici producendone analisi più precise e attendibili.

Appendice

```
#Intervallo di confidenza alla WALD
p<-x/n
n<-length(dati)
p
alfa<-0.05
zalfa<-qnorm(1-alfa/2)
zalfa
ci<-c(p-zalfa*sqrt(p*(1-p)/n), p+zalfa*sqrt(p*(1-p)/n) )
ci

#Intervallo di confidenza score
alfa<-0.05
zalfa<-qnorm(1-alfa/2)
zalfa
ci<-c(
(p+((zalfa^2)/(2*n))-zalfa*sqrt((p*(1-p)+(zalfa^2)/(4*n))/n))
/(1+(zalfa^2)/n)
,(p+((zalfa^2)/(2*n))+zalfa*sqrt((p*(1-p)+(zalfa^2)/(4*n))/n))
/(1+(zalfa^2)/n)
)
ci

#Intervallo di confidenza alla WALD AGGIUSTATO
x1<-x+2
```

```
n1<-n+4
p<-x1/n1
p
alfa<-0.05
zalfa<-qnorm(1-alfa/2)
zalfa
ci<-c(p-zalfa*sqrt(p*(1-p)/n1), p+zalfa*sqrt(p*(1-p)/n1) )
ci

#INFERENZA CONDIZIONATA APPROSSIMATA PER I DATI

tab1<-data.frame(n=c(6.5,0.5),t=c(10,1),x=c(0,1))
tab1
summary(tab1.fit<-glm (cbind(n,t-n)~x,binomial,data=tab1))
summary(cond.glm(tab1.fit,offset=x),test=0)
```

Riferimenti bibliografici

Agresti A., Coull B. A. (1998). Approximate is Better than “Exact” for Interval Estimation of Binomial Proportions. *The American Statistician*, **52**, 119-126.

Azzalini A. (1992). *Inferenza Statistica: Un’introduzione Basata sul Concetto di Verosimiglianza*. Springer-Verlag, Heidelberg.

Azzalini A. (2000). *Inferenza Statistica: una presentazione basata sul concetto di verosimiglianza*. Springer, Milano.

Barndorff-Nielsen O.E. (1980). Conditionality resolutions. *Biometrika*, **67**, 293-310.

Barndorff-Nielsen O.E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343-365.

Barndorff-Nielsen O.E. (1991b). Modified signed log likelihood ratio. *Biometrika*, **78**, 557-563.

Besana R., Boroli A., Drago M. (1995). *Enciclopedia della medicina*. Istituto Geografico De agostini, Novara.

Bortot P., Salvan A., Ventura L. (2000). *Inferenza Statistica: Applicazioni con S-PLUS e R.* CEDAM, Padova.

Brazzale A.R., Davison A.C. (2009). *Accurate Parametric Inference for Small Sampler*. Statistical Science, in corso di pubblicazione.

Brazzale A.R., Davison A.C., Reid N. (2007). *Applied Asympotics*. Cambrige University Press, Cambridge.

Brown L.D., Cai T.T., DasGupta A. (2001). *Interval Estimation for a Binomial Proportion*, **16**, 101-133.

Cox D.R., Reid N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc.* B, **49**, 1-39.

Iacus S. M., Masarotto G. (2003). *Laboratorio di statistica con R.* McGraw-Hill, Milano.

Moroni M., Esposito R., De Lalla F. (2002). *Malattie infettive*. Elsevier, Milano.

Murray P. R., Rosenthal K. S., Kobayashi G. S. (2003). *Microbiologia*. Edises s.r.l., Napoli.

Newcombe R.G. (1998). Two-sided confidence intervals for the single proportion; comparison of several methods, *Statistics in Medicine*, **17**, 857-872.

Pace L., Salvan A. (1996). *Teoria della Statistica. Metodi, modelli, approssimazioni asintotiche.* CEDAM, Padova.

Pace L., Salvan A. (2001). *Introduzione alla statistica. Inferenza, verosimiglianza, modelli.* CEDAM, Padova.

Piccolo D. (1998). *Statistica*. Il Mulino, Bologna.

Severini T. A. (2000). *Likelihood Methods in Statistics*. Oxford University Press, Oxford.

Vollset S.E. (1993). Confidence intervals for a binomial proportion, *Statistics in Medicine*, **12**, 809-824.

Siti Internet utili

Dal sito internet della giunta regionale (2007). *Linee guida per il controllo della tubercolosi nella Regione Veneto*. Materiale disponibile on line, http://bur.regione.veneto.it/BurvServices/pubblica/DettaglioDgr.aspx?id=198620.

E' possibile scaricare gratuitamente il software R. dal sito internet, http://www.r-project.org/.