

Premessa

*Dove è la sapienza
che abbiamo smarrita nella conoscenza
dove è la conoscenza
che abbiamo smarrita nell'informazione
dove è l'informazione
che abbiamo smarrita nei dati.*

(Mark Porat)

In tutte le scienze della vita, e nell'ambito biomedico in particolare, le analisi statistiche di fenomeni reali comportano sia la conoscenza dei metodi appropriati sia l'uso del calcolo automatico che permette la manipolazione di dati a volte complessi, il cui trattamento sarebbe proibitivo senza l'ausilio di un *software* adatto.

Questo volume si propone di divulgare la metodologia statistica per la ricerca nelle scienze della vita, attraverso l'insegnamento dei principali metodi statistici applicati a casi di studio, con l'ausilio del calcolo automatico. Ogni capitolo è stato impostato, dunque, a partire da un caso di studio (o più d'uno) presentato come problema motivante per l'introduzione delle basi teoriche statistiche su cui poggiare gli aspetti più tecnici necessari allo studio e/o alla risoluzione dei problemi. Una particolare cura è posta nel mostrare il percorso che conduce alla formalizzazione del problema in termini logici e simbolici e, successivamente, sono illustrate le metodologie statistico-matematiche e gli strumenti tecnici che possono portare alla sua risoluzione.

Poiché il volume intende rivolgersi agli statistici ma anche ai ricercatori nelle scienze della vita e a un'ampia platea di studenti di diversi corsi di laurea (biologia, medicina, statistica medica, biotecnologie, farmacia, scienze, agraria, ecc.), l'impostazione espositiva consente due letture tra loro complementari. Da un lato il libro può essere visto come una raccolta ragionata e concatenata di tecniche per l'analisi statistica di dati sperimentali dove, partendo da di-

versi casi reali, sono illustrate sia le domande che il fenomeno in studio pone all'esperto sia i metodi che possono condurre alle risposte più appropriate, con l'ausilio del calcolo automatico. D'altro lato può essere riguardato come un manuale d'uso del *software* R, certamente il più diffuso nell'ambiente scientifico, trattandosi di un *software open-source* che contiene una immensa libreria di programmi pronti o facilmente adattabili a ogni tipo di problema statistico-probabilistico. Il lettore che già possiede le nozioni di statistica di base, può dunque utilizzarlo per apprendere, attraverso esempi, il linguaggio R applicato alla ricerca in biomedicina.

Per entrambe le tipologie d'uso, il libro offre anche alcuni esempi di problemi che richiedono l'applicazione di tecniche statistiche più avanzate e di comandi in R di non immediata intuizione, rimandando alla bibliografia per gli approfondimenti sia teorici sia tecnici.

La stesura di questo volume nasce dall'esperienza degli autori che, attraverso collaborazioni con numerosi studiosi nei diversi campi della biomedicina e le esperienze didattiche acquisite nei corsi di Statistica, Modelli Statistici e Statistica Medica, hanno maturato la convinzione che l'esperto biomedico non debba sostituirsi allo statistico, "distraendosi" dai propri studi specifici. Ma è certamente indispensabile che egli si renda autonomo per quanto riguarda le nozioni e le applicazioni di base della statistica, in modo da poter effettuare le analisi statistiche che devono costituire il corredo culturale irrinunciabile del ricercatore e anche affinché, nell'interazione con lo statistico per i problemi di ricerca più avanzati, sia in grado di interagire con competenza. È infatti importante che l'esperto di dominio acquisisca conoscenza delle procedure di analisi statistica per poter ampliare le richieste da porre ai dati - e dunque allo statistico - e sviluppare la capacità interpretativa delle restituzioni che il calcolo automatico fornisce. In tale modo i frutti degli studi su problemi reali possono condurre a ricerche più avanzate nei due ambiti: biomedico e statistico.

Per coloro che intraprendono lo studio della Statistica come disciplina privilegiata della propria formazione, il volume vuole essere una sorta di *vademecum* che, oltre a percorrere i metodi più diffusi dell'analisi statistica, entra nel merito dell'elaborazione di numerose applicazioni, con particolare riguardo alla scelta delle tecniche più appropriate e all'uso dell'ambiente R.

Si presuppone che il lettore abbia già una discreta preparazione sui concetti presentati in un corso di statistica di base e, possibilmente, dell'ambiente R. Alcuni testi utili a questo scopo sono elencati nei riferimenti bibliografici¹. Una

¹Si segnalano in particolare Agresti e Finlay (2009), Bland (2009), Piccinato (2009), Triola e Triola (2009), Wayne (2009), Piccolo (2010), Cicchitelli (2012) e Corradi (2014).

copia di R può essere scaricata all'indirizzo *web* `cran.r-project.org/bin`, dove si trovano versioni del linguaggio per diversi sistemi operativi. Manuali introduttivi all'uso di R sono recuperabili all'indirizzo `cran.r-project.org`². Diversi capitoli, in particolare dove sono trattati argomenti che richiedono nozioni statistiche non propriamente “di base”, sono completati da un paragrafo dedicato ai richiami teorici di quelle nozioni. Alcuni di questi sono facilmente affrontabili anche a partire da conoscenze elementari, in altri casi è necessaria una preparazione più approfondita. D'altra parte sarebbe impossibile raccogliere in un solo volume tutti gli sviluppi che la Statistica moderna (ma non solo!) ha avuto. Si è cercato di sopperire indicando i rimandi bibliografici nelle numerose note che corredano quasi ogni pagina.

Lo schema del volume è il seguente. Esso si apre con l'introduzione degli strumenti classici dell'analisi esplorativa e descrittiva, per arrivare ai test di adattamento. I Capp. 3-6 presentano i principali test parametrici e non parametrici, a uno o più campioni indipendenti, per dati appaiati e misure ripetute e per l'analisi di tabelle di contingenza. Il Cap. 7 è focalizzato sull'accuratezza dei test diagnostici, illustrando la curva ROC e gli strumenti a essa associati. I Capp. 8-10 sono dedicati ai principali modelli di regressione utilizzati nelle analisi biomediche e in altri campi di ricerca, (regressione lineare, regressione logistica e GLM, modelli dose-risposta). Infine, nel Cap. 11 è trattata l'analisi della sopravvivenza non parametrica e semi-parametrica. A conclusione di ogni capitolo, sono proposti alcuni *dataset* con i quali esercitarsi. Chiude il volume una Appendice che richiama sinteticamente i tipi di studio e la determinazione della numerosità campionaria.

Il contenuto del volume può anche essere pensato suddiviso in due moduli di diversa difficoltà. Un primo modulo, basato sui primi capitoli (Capp. 1-6), fa riferimento ad argomenti usualmente presentati in corsi di base di diversi corsi di laurea. Un secondo modulo, basato sui successivi capitoli (Capp. 7-11), introduce i principali modelli di regressione, che richiedono alcune conoscenze più dettagliate dell'inferenza. La Fig. 1 riporta, in maniera semplificata, un riassunto dei metodi utilizzati nel volume.

²Particolarmente indicati sono Muggeo e Ferrara (2005) e Maindonald (2008), oltre ai volumi Bortot *et al.* (2000) e Iacus e Masarotto (2013).

Questo libro contiene almeno un errore. Ci si potrebbe aspettare che per verificare la cosa sia necessario leggere l'intero volume. E invece lo sappiamo già fin d'ora. Infatti, se ci sono errori, ci sono. E se non ce ne sono, c'è quello che dice: "Questo libro contiene almeno un errore". Dunque sappiamo che in questo libro un errore c'è, anche se non sappiamo ancora qual è.

(P. Odifreddi)

La responsabilità di ogni errore e imprecisione presenti in questo volume è da attribuirsi esclusivamente agli autori, che saranno grati a chi li segnalerà. Il manoscritto è stato redatto in L^AT_EX dagli stessi autori. Nel sito

<http://mybook.egeaonline.it>

oltre a una versione aggiornata dell'*errata corrige*, è offerta al lettore la possibilità di recuperare e utilizzare gli insiemi di dati, tipicamente di piccole dimensioni, usati nei casi di studio. Critiche, commenti e segnalazioni possono essere inviate agli autori agli indirizzi email ventura@stat.unipd.it e racugno@unica.it.

Laura e Walter