

HARVEY MOTULSKY

# Biostatistica essenziale

*Una guida non matematica*



*Edizione italiana a cura di*

**Leonardo Emberti Gialloreti**

Dipartimento di Biomedicina e Prevenzione

Università degli Studi di Roma "Tor Vergata"

*Traduzione di*

**Daniele di Giovanni**

Dipartimento di Ingegneria Industriale

Università degli Studi di Roma "Tor Vergata"

*Con la collaborazione di*

**Fabio Riccardi**

**Annalisa Aragozzini**

**PICCIN**

This is a translation of  
*ESSENTIAL BIOSTATISTICS: A Nonmathematical Approach*, First Edition  
by Harvey Motulsky  
Copyright © 2016 by Oxford University Press

*Essential Biostatistics: A Nonmathematical Approach*, First Edition was originally published in English in 2016. This translation is published by arrangement with Oxford University Press. Piccin Nuova Libreria is solely responsible for this translation from the original work and Oxford University Press shall have no liability for any errors, omissions or inaccuracies or ambiguities in such translation or for any losses caused by reliance thereon.

L'edizione originale in lingua inglese di *Essential Biostatistics: A Nonmathematical Approach*, First Edition è stata pubblicata nel 2016. Questa traduzione è stata pubblicata su licenza di Oxford University Press. Piccin Nuova Libreria è responsabile della traduzione dall'opera originale e Oxford University Press non è responsabile per eventuali errori, omissioni, sviste o ambiguità della traduzione o per eventuali danni da essa derivanti.

Opera coperta dal diritto d'autore – tutti i diritti sono riservati.  
Questo testo contiene materiale, testi ed immagini, coperto da copyright e non può essere copiato, riprodotto, distribuito, trasferito, noleggiato, licenziato o trasmesso in pubblico, venduto, prestato a terzi, in tutto o in parte, o utilizzato in alcun altro modo o altrimenti diffuso, se non previa espressa autorizzazione dell'editore. Qualsiasi distribuzione o fruizione non autorizzata del presente testo, così come l'alterazione delle informazioni elettroniche, costituisce una violazione dei diritti dell'editore e dell'autore e sarà sanzionata civilmente e penalmente secondo quanto previsto dalla L. 633/1941 e ss.mm.

ISBN 978-88-299-3162-0

Stampato in Italia

# INDICE GENERALE



PRESENTAZIONE ALL'EDIZIONE ITALIANA XI

PREFAZIONE XIII

## **1. Statistica e probabilità non sono concetti intuitivi 1**

Tendiamo a saltare subito alle conclusioni 1

Tendiamo a essere troppo sicuri di noi stessi 1

Vediamo tendenze nei dati casuali 2

Non ci aspettiamo che la variabilità dipenda dalla dimensione del campione 3

I confronti multipli ci ingannano 4

Tendiamo a ignorare possibili spiegazioni alternative 5

Desidereremmo conclusioni univoche ma la statistica ci propone delle probabilità 6

Sintesi del capitolo 6

## **2. La probabilità è un concetto complesso 7**

Elementi base di probabilità 7

Probabilità come previsione di frequenza sul lungo periodo 7

Probabilità come forza della convinzione personale (Bayes) 8

La differenza tra probabilità e statistica 9

Terminologia 10

Errori frequenti 11

Sintesi del capitolo 12

## **3. Dal campione alla popolazione 14**

Estrarre un campione da una popolazione 14

Fino a che punto è possibile generalizzare? 15

Terminologia 15

Errori frequenti 17

Sintesi del capitolo 18

**4. Gli intervalli di confidenza 19**

- Esempio: sopravvivenza di neonati prematuri 19
- Esempio: sondaggio elettorale 20
- Presupposti: l'intervallo di confidenza di una proporzione 21
- Cosa significa veramente "confidenza al 95%"? 22
- Gli intervalli di confidenza vanno interpretati nel loro contesto 25
- Intervalli di confidenza per altre tipologie di dati 27
- Terminologia 29
- Errori frequenti 31
- Domande e risposte 31
- Sintesi del capitolo 33

**5. Tipi di variabili 34**

- Variabili continue 34
- Variabili ordinali 36
- Variabili nominali 36
- Domande e risposte 36
- Sintesi del capitolo 37

**6. Rappresentazione grafica della variabilità 39**

- Rappresentazione grafica della dispersione e della distribuzione dei dati 39
- Attenzione ai dati preprocessati 43
- Terminologia 45
- Errori frequenti 46
- Domande e risposte 47
- Sintesi del capitolo 47

**7. Quantificare la variabilità 48**

- Intervallo (*range*) 48
- Percentili 48
- Intervallo interquartile 48
- Cinque numeri di sintesi 49
- Deviazione standard 49
- Coefficiente di variazione 51
- Terminologia 52
- Errori frequenti 52
- Domande e risposte 53
- Sintesi del capitolo 53

- 8. La distribuzione gaussiana 55**  
Da dove deriva la distribuzione gaussiana? 55  
Il significato della deviazione standard in una distribuzione gaussiana 56  
Come appare realmente un campione estratto da una distribuzione gaussiana? 57  
Perché la distribuzione gaussiana è così essenziale in statistica 57  
Terminologia 59  
Errori frequenti 59  
Domande e risposte 60  
Sintesi del capitolo 60
- 9. La distribuzione log-normale e la media geometrica 61**  
Panoramica 61  
Esempio: rilassare le vesciche 61  
Rivediamo i logaritmi 61  
L'origine della distribuzione log-normale 63  
Come analizzare dati log-normali 64  
Media geometrica 64  
Terminologia 65  
Errori frequenti 65  
Domande e risposte 65  
Sintesi del capitolo 66
- 10. L'intervallo di confidenza per la media 67**  
Interpretare l'intervallo di confidenza per la media 67  
Quali valori determinano l'intervallo di confidenza per la media? 68  
L'errore standard della media 69  
Presupposti dell'intervallo di confidenza per la media 70  
Terminologia 71  
Errori frequenti 72  
Domande e risposte 72  
Sintesi del capitolo 74
- 11. Le barre di errore 75**  
Le diverse tipologie di barre di errore 75  
Come interpretare le barre di errore 76  
Che tipo di barra di errore utilizzare? 77

In che modo la deviazione standard e l'errore standard della media sono collegati alle dimensioni del campione? 78

Terminologia 79

Errori frequenti 79

Domande e risposte 81

Sintesi del capitolo 82

## **12. Confronto tra gruppi attraverso gli intervalli di confidenza 83**

Utilizzare gli intervalli di confidenza per confrontare diversi gruppi di variabili 83

Alcuni esempi di intervalli di confidenza utilizzati per confrontare i gruppi 83

Presupposti degli intervalli di confidenza 90

Errori frequenti 90

Domande e risposte 91

Sintesi del capitolo 91

## **13. Confronto tra gruppi attraverso il $p$ -value 92**

Introduzione al  $p$ -value attraverso "testa o croce" 92

Una regola che collega l'intervallo di confidenza con il  $p$ -value 93

Riesame degli esempi del Capitolo 12 94

Quattro cose che dovete sapere sul  $p$ -value 97

Terminologia 98

Errori frequenti 98

Domande e risposte 101

Sintesi del capitolo 102

## **14. I test di significatività statistica e i test di ipotesi 103**

Testare un'ipotesi attraverso la statistica 103

Riesame degli esempi dei Capitoli 12 e 13 103

Un'analogia: innocente fino a prova contraria 104

Estremamente significativo? al limite della significatività? 105

Gli errori di tipo I, II e III 106

Scegliere il livello di significatività 108

Terminologia 110

Errori frequenti 110

Domande e risposte 112

Sintesi del capitolo 112

- 15. Interpretare un risultato che è (o non è) statisticamente significativo** 114
- Interpretare risultati “statisticamente significativi” 114
  - Interpretare risultati “non statisticamente significativi” 117
  - Cinque possibili motivi per cui si è ottenuto un risultato “non statisticamente significativo” 119
  - Terminologia 120
  - Errori frequenti 121
  - Domande e risposte 121
  - Sintesi del capitolo 121
- 16. Quanto sono comuni gli errori di tipo I?** 123
- Cos’è un errore di tipo I? 123
  - Con quale frequenza si verificano gli errori di tipo I? 123
  - La probabilità a priori influenza il *false discovery rate* (un po’ di Bayes) 125
  - Analogia con i test clinici 128
  - Terminologia 129
  - Errori frequenti 130
  - Domande e risposte 130
  - Sintesi del capitolo 131
- 17. I confronti multipli** 132
- Perché i confronti multipli sono un problema? 132
  - Un esempio di problema con i confronti multipli 132
  - Confronti multipli in diversi contesti 134
  - Come affrontare i confronti multipli 137
  - Terminologia 140
  - Errori frequenti 140
  - Domande e risposte 140
  - Sintesi del capitolo 141
- 18. La potenza statistica e la dimensione del campione** 142
- Determinare la dimensione di un campione nel corso di uno studio, e non prima di iniziarlo, porta a risultati fuorvianti 142
  - Quattro domande sulla dimensione del campione 143
  - Come si legge la descrizione della dimensione del campione 144
  - La dimensione di un campione: un calcolo o una trattativa? 145

Un'analogia per comprendere la potenza statistica	146
Dimensione del campione e margine di errore dell'intervallo di confidenza	147
Terminologia	147
Errori frequenti	147
Domande e risposte	149
Sintesi del capitolo	150

**19. Test statistici di uso comune** 151

Presupposti di tutti i test statistici di uso comune	151
Confronto di una variabile continua tra due gruppi	152
Confronto di una variabile continua tra tre o più gruppi	155
Confronto di una variabile binaria (dicotomica) tra due gruppi	157
Confronto delle curve di sopravvivenza	159
Correlazione e regressione	159
Terminologia	159
Sintesi del capitolo	159

**20. I test di normalità** 160

Valutare la normalità di una distribuzione	160
Problemi connessi all'uso dei test di normalità	160
Alternative al presupposto della distribuzione gaussiana	161
Terminologia	162
Errori frequenti	162
Domande e risposte	163
Sintesi del capitolo	163

**21. Gli outlier** 164

Da dove derivano gli <i>outlier</i> ?	164
Test per gli <i>outlier</i>	164
Cinque domande da porsi prima di effettuare un test per gli <i>outlier</i>	165
La domanda a cui risponde un test per gli <i>outlier</i>	166
È legittimo rimuovere gli <i>outlier</i> ?	167
Terminologia	167
Errori frequenti	167
Domande e risposte	169
Sintesi del capitolo	169

- 22. La correlazione** 171  
 Introduzione al coefficiente di correlazione 171  
 Presupposti della correlazione 175  
 Terminologia 176  
 Errori frequenti 177  
 Domande e risposte 178  
 Sintesi del capitolo 179
- 23. La regressione lineare semplice** 180  
 A cosa serve la regressione lineare 180  
 Risultati della regressione lineare 180  
 Presupposti della regressione lineare 184  
 Confronto tra regressione lineare e correlazione 186  
 Terminologia 186  
 Errori frequenti 187  
 Domande e risposte 192  
 Sintesi del capitolo 193
- 24. Regressione non lineare, regressione multipla e regressione logistica** 194  
 Regressione non lineare 194  
 Regressione multipla e regressione logistica 195  
 Terminologia 196  
 Errori frequenti 196  
 Domande e risposte 198  
 Sintesi del capitolo 198
- 25. Errori da evitare** 200  
 Errore: non riconoscere eventuali *bias* di pubblicazione 200  
 Errore: verificare ipotesi suggerite dai dati e non formulate prima 201  
 Errore: vedere una causalità ogni volta che i dati indicano una correlazione 201  
 Errore: sopravvalutare gli studi che analizzano variabili *proxy* (surrogate) 201  
 Errore: sopravvalutare i risultati di uno studio osservazionale 203  
 Errore: farsi ingannare dalla “regressione verso la media” 204

<b>26. Sintesi conclusiva</b>	206
I concetti fondamentali della statistica	206
Termini statistici utilizzati nei capitoli	207
<b>BIBLIOGRAFIA</b>	211
<b>INDICE ANALITICO</b>	217

## PRESENTAZIONE ALL'EDIZIONE ITALIANA



Comprendere a fondo un lavoro scientifico e fare ricerca è impossibile senza una qualche nozione di statistica. Oggi fare un'analisi statistica è molto più facile che nel passato per la semplicità di utilizzo dei molti software in commercio. L'accessibilità a questi strumenti rende più semplice, per tutti, affrontare l'analisi dei dati ma di pari passo aumenta il rischio di commettere errori e di giungere a conclusioni sbagliate. Questo accade molto più spesso di quanto si pensi.

Si può *imparare* la statistica su molti eccellenti testi. Il volume che avete in mano è invece pensato per chi vuole *imparare a utilizzare* la statistica. Imparare e imparare ad utilizzare: posso dire, dopo tanti anni di insegnamento universitario, che non è una differenza da poco. Molti infatti possono fare un calcolo statistico, ma non tutti sono in grado di affrontare una situazione reale che richiede di applicare un'analisi statistica e di comprenderne il senso.

È per questo motivo che volentieri ho accolto l'invito dell'Editore Piccin di curare *Biostatistica essenziale: una guida non matematica*, edizione italiana del rinomato *Essential Biostatistics: A Nonmathematical Approach* di Harvey Motulsky, un testo che è progettato per chi vuole concretamente utilizzare la statistica in prima persona, ma che è pensato anche per chi desidera comprendere il reale messaggio che i dati statistici possono trasmettere.

È un testo da utilizzare come compagno di viaggio, in parallelo o quale approfondimento ad un corso universitario di statistica sanitaria. Non si sofferma su tutti i dettagli (qualche purista forse storcerà il naso, pazienza) ma accompagna il lettore a *fare statistica* e ad accorgersi in tempo dei possibili *errori* in cui è facile incorrere.

Speriamo di offrire alla comunità universitaria e ai ricercatori italiani un agevole *sussidio* per scegliere con competenza le tecniche statistiche più utili a rispondere a domande scientifiche reali e per comprendere e utilizzare consapevolmente la terminologia statistica.

*Prof. Leonardo Emberti Gialloreti*



## PREFAZIONE



*Biostatistica essenziale: una guida non matematica* è un sussidio non matematico, conciso ed accessibile, al pensiero statistico. È concepita come un breve testo di accompagnamento ai corsi di statistica universitaria e alla lettura di libri di statistica più lunghi che adottano un approccio più matematico o anche come strumento di revisione per i ricercatori.

### **In che modo questa edizione è diversa dall'edizione completa di *Intuitive Biostatistics*?**

*Intuitive Biostatistics* venne pubblicato per la prima volta nel 1995, ed è ora alla sua quarta edizione. Tuttavia, sebbene in molti abbiano elogiato il libro, alcuni docenti lo trovavano poco adatto al loro corso perché non comprendeva tutti quei contenuti matematici che ci si attenderebbe di trovare in un libro di testo principale; d'altra parte, risultava anche troppo lungo per essere utilizzato come testo complementare. Per colmare questa lacuna, ho scritto *Biostatistica essenziale: una guida non matematica*, lungo circa un terzo di *Intuitive Biostatistics* (3ª edizione).

Alcuni tra gli argomenti presenti in *Intuitive Biostatistics* ma assenti da *Biostatistica essenziale: una guida non matematica* sono la meta-analisi, le curve di sopravvivenza, i test di equivalenza o di non inferiorità, la regressione non lineare e multipla, il confronto della bontà di adattamento tra modelli alternativi e o le curve ROC. Inoltre, *Intuitive Biostatistics* (ma non *Biostatistica essenziale: una guida non matematica*) comprendeva anche più di 100 pagine di spiegazione dei test statistici di base, nonché 40 pagine di esercizi con risposte.

### **Un approccio unico**

Ecco alcuni aspetti che rendono questo libro unico nel suo genere:

- Il Capitolo 1 spiega come il senso comune possa portare a conclusioni errate e perché è necessario comprendere i principi della statistica.
- Il Capitolo 2 propone un approccio unico per rendersi conto di quanto sia complesso il concetto di probabilità.

- Introduco il pensiero statistico a partire dal Capitolo 4, dove si spiega l'intervallo di confidenza per la proporzione. Posso così spiegare la logica alla base dell'utilizzo dell'intervallo di confidenza per generalizzare da un campione ad una popolazione, prima di dover affrontare il concetto della dispersione. Non commettete l'errore di trascurare il Capitolo 4 solo perché pensate di non dover lavorare con dati espressi come delle proporzioni, poiché questo capitolo spiega alcuni dei più importanti concetti della statistica.
- Introduco il confronto tra gruppi tramite gli intervalli di confidenza (Capitolo 12) prima di spiegare il *p-value* (Capitolo 13) e la significatività statistica (Capitoli 14 e 15). In questo modo posticipo la presentazione dei concetti disorientanti e fin troppo abusati di *p-value* e di "significativo".
- Il Capitolo 16 spiega quanto sono comuni gli errori di tipo I e qual è la differenza tra significatività e *false discovery rate*.
- Il Capitolo 19 presenta, sotto forma di tabelle, i più comuni test statistici:
- Ho incluso alcuni argomenti spesso omissi dai testi introduttivi, ma che considero essenziali. Ad esempio, i confronti multipli, il *false discovery rate*, il "dragaggio dei dati" (*p-hacking*), le distribuzioni log-normali, la media geometrica, i test di normalità, gli *outlier* o la regressione non lineare.
- Quasi ogni capitolo ha una sezione dedicata alla *terminologia*, dove si spiegano alcuni possibili fraintendimenti linguistici.
- Per aiutarvi ad evitare interpretazioni errate dei dati, quasi ogni capitolo include anche una sezione che discute gli *errori più frequenti*. L'intero Capitolo 25 è, infine, incentrato sui principali errori da evitare.

### Chi ha aiutato?

Un immenso grazie alle molte persone che hanno rivisto le bozze dei capitoli. I loro contributi hanno enormemente migliorato questo libro:

Abdel-Salam G. Abdel-Salam, Virginia Polytechnic Institute and State University

B. Carol Adjemian, Pepperdine University

Raid Amin, University of West Florida

Michael Biro, Swarthmore College

Patrick Breheny, University of Kentucky

Michael F. Cassidy, Marymount University

Dean W. Coble, Stephen F. Austin State University

William M. Cook, Saint Cloud State University

Erica A. Corbett, Southeastern Oklahoma State University

Vincent A. DeBari, Seton Hall University

Bianca Frogner, George Washington University

Robert M. Hamer, University of North Carolina at Chapel Hill

Philip Hejduk, University of Texas at Arlington

Ravi P. Joshi, Old Dominion University  
 Stefan Judex, Stony Brook University  
 Chris Kerth, Texas A&M University  
 Joshua Lallaman, Saint Mary's University of Minnesota  
 Gary A. Lamberti, University of Notre Dame  
 Bruce D. Leopold, Mississippi State University  
 Susan P. McGorray, University of Florida  
 Matt McQueen, University of Colorado Boulder  
 Sumona Mondal, Clarkson University  
 Christopher J. Salice, Texas Tech University  
 David A. Sanchez, University of Texas at Arlington  
 Evelyn H. Schlenker, University of South Dakota  
 Andrew Jay Tierman, Saginaw Valley State University  
 Kathryn Trinkaus, Washington University  
 Derek Webb, Bemidji State University  
 Mary M. Whiteside, University of Texas at Arlington  
 John W. Wilson, University of Pittsburgh  
 Naji Younes, George Washington University

Ringrazio anche tutte le persone della Oxford University Press che hanno contribuito a trasformare il mio manoscritto in un libro raffinato: Jason Noe, Senior Editor; Andrew Heaton, Assistant Editor; Patrick Lynch, Editorial Director; John Challice, Publisher and Vice President; Bill Marting, National Sales Manager; Frank Mortimer, Director of Marketing; David Jurman, Marketing Manager; Elizabeth Geist, Marketing Assistant; Lisa Grzan, Production Manager; Amy Whitmer, Production Team Leader; Christian Holdener, Senior Production Manager; Michele Laseau, Art Director; e Bonni Leon-Berman, Designer.

### **La mia esperienza in materia**

Per oltre un decennio ho insegnato statistica agli studenti del primo anno di medicina e ai laureati in scienze biomediche dell'Università della California, San Diego. Il programma di questi corsi si è ampliato fino a divenire la prima edizione di *Intuitive Biostatistics*, il testo originale e completo su cui si basa questo libro. Non insegno più in questi corsi, ma in qualità di fondatore e CEO di GraphPad Software, scambio e-mail con studenti e ricercatori quasi ogni giorno. Resto quindi costantemente aggiornato sui molti modi in cui i concetti statistici possono confondere o essere fraintesi.

Scrivetemi per consigli, correzioni o suggerimenti. Eventuali correzioni saranno riportate nel sito [www.intuitivebiostatistics.com](http://www.intuitivebiostatistics.com).

Harvey Motulsky  
[hmotulsky@graphpad.com](mailto:hmotulsky@graphpad.com)

