Sommario

Prefazione **ix**

1 Un campione **1**
1.1 Introduzione . . . 1
1.1.1 Esempi . . . 1
1.1.2 Analogie e differenze . . . 5
1.2 Una media . . . 8
1.2.1 La verifica di ipotesi su di una media: l'approccio di Fisher del p-value . . . 10
1.2.2 La verifica di ipotesi su di una media: l'approccio di Neyman e Pearson della regione critica . . . 13
1.2.3 p-value e livello di significatività . . . 22
1.2.4 Il test z per una media . . . 24
1.2.5 Il test t di Student per una media . . . 26
1.2.6 Intervallo di confidenza per una media (σ noto) . . . 31
1.2.7 Intervallo di confidenza e test di significatività . . . 36
1.2.8 Intervallo di confidenza per una media (σ non noto) . . . 37
1.3 Metodi non parametrici . . . 39
1.3.1 Il grafico quantile-quantile per una distribuzione gaussiana . . . 40
1.3.2 Il test di Kolmogorov-Smirnov . . . 41
1.3.3 Il test di Shapiro-Wilk . . . 44
1.4 Metodi di ricampionamento . . . 46
1.4.1 Test di permutazione . . . 46
1.4.2 Il bootstrap . . . 48
1.5 Una proporzione . . . 52
1.5.1 Il test z per una proporzione . . . 53
1.5.2 Intervallo di confidenza (approssimato) per una proporzione . . . 56
1.5.3 Test "esatto" per una proporzione . . . 59
1.5.4 Intervallo di confidenza "esatto" per una proporzione . . . 62
1.5.5 Intervalli di confidenza per una proporzione: meglio "esatti" o approssimati? . . . 64
1.6 Un conteggio . . . 69
1.6.1 Test e intervalli di confidenza approssimati . . . 69
1.6.2 Il test di dispersione per una Poisson . . . 73
1.6.3 Test e intervalli di confidenza "esatti" . . . 75

2 Due campioni indipendenti 79
2.1 Introduzione . . . 79
2.1.1 Esempi . . . 79
2.1.2 Analogie e differenze . . . 82
2.2 Due medie . . . 84
2.2.1 Il test t di Student per due campioni indipendenti . . . 84
2.2.2 Intervallo di confidenza per la differenza fra le medie di due campioni indipendenti . . . 89
2.2.3 Il caso di varianze diverse: il test di Welch . . . 90
2.3 Metodi non parametrici . . . 91
2.3.1 Il test di Wilcoxon per due campioni indipendenti . . . 92
2.3.2 Il test di Kolmogorov-Smirnov per due campioni . . . 94
2.3.3 Il test delle "successioni" di Wald e Wolfowitz . . . 98
2.4 Metodi di ricampionamento . . . 101
2.4.1 Test di permutazione . . . 101
2.4.2 Il bootstrap . . . 108
2.5 Due proporzioni . . . 111
2.5.1 Il test z per il confronto di due proporzioni . . . 111
2.5.2 Intervallo di confidenza per la differenza fra due proporzioni . 113
2.5.3 Il test chi quadrato per il confronto di due proporzioni . . . 114
2.5.4 Il test esatto di Fisher . . . 119
2.5.5 Inferenza sull'odds ratio . . . 124
2.6 Due conteggi . . . 129

3 Più di due campioni indipendenti 137
3.1 Introduzione . . . 137
3.1.1 Esempi . . . 137
3.1.2 Analogie e differenze . . . 142
3.2 Medie . . . 142
3.2.1 Test *t* con la correzione di Bonferroni . . . 142
3.2.2 L'analisi della varianza . . . 145
3.2.3 Confronti multipli . . . 156
3.3 Metodi non parametrici . . . 177
3.3.1 L'analisi della varianza di Kruskal–Wallis . . . 177
3.4 Metodi di ricampionamento . . . 180
3.5 Proporzioni . . . 184
3.5.1 Il test chi quadrato per il confronto di più di due proporzioni . 184
3.5.2 Confronti multipli tra proporzioni . . . 187
3.5.3 Il chi-quadrato per il trend . . . 194
3.5.4 Il test esatto di Fisher . . . 196
3.6 Conteggi . . . 197

4 Campioni dipendenti **201**
4.1 Introduzione . . . 201
4.1.1 Esempi . . . 201
4.1.2 Analogie e differenze . . . 205
4.2 Medie . . . 205
4.2.1 Il test t di Student per due campioni dipendenti . . . 205
4.2.2 Intervallo di confidenza per la differenza fra le medie di due campioni dipendenti . . . 209
4.2.3 L'analisi della varianza per campioni dipendenti . . . 211
4.2.4 La "sfericità" . . . 214
4.3 Metodi non parametrici . . . 218
4.3.1 Il test dei segni . . . 218
4.3.2 Il test di Wilcoxon per due campioni dipendenti . . . 220
4.3.3 Il test di Friedman per più di due campioni dipendenti . . . 223
4.4 Metodi di ricampionamento . . . 225
4.5 Proporzioni . . . 228
4.5.1 Confronto di due proporzioni dipendenti . . . 228
4.5.2 Intervallo di confidenza per la differenza di due proporzioni dipendenti . . . 234
4.5.3 Intervallo di confidenza per l'*odds ratio* nel caso di due campioni dipendenti . . . 235
4.5.4 Confronto di più di due proporzioni: il test Q di Cochran . . . 236
4.6 Conteggi . . . 240

5 Misure di effetto, dimensione del campione, potenza **243**
5.1 Alcune considerazioni sui test di significatività . . . 243
5.2 Misure di effetto . . . 248
5.3 Dimensione del campione e ampiezza dell'intervallo di confidenza . . . 250
5.3.1 Una media . . . 250
5.3.2 Una proporzione . . . 253
5.4 Dimensione del campione e potenza di un test . . . 257
5.4.1 Una media . . . 257
5.4.2 Una proporzione . . . 263
5.5 Il calcolo della potenza e della dimensione del campione con R . . . 267
5.5.1 Test *t* per una e per due medie . . . 267
5.5.2 Test per due proporzioni . . . 271
5.5.3 Analisi della varianza (ANOVA) . . . 273
5.6 "Dimostrare" l'ipotesi nulla: i test di equivalenza . . . 274

Bibliografia **277**

Indice analitico **281**

Indice dei termini di R **285**

Indice dei nomi **287**

Prefazione

Il libro che state sfogliando avrebbe dovuto intitolarsi R *per la* R*icerca*. Ragioni per lo più editoriali hanno portato a preferire il titolo attuale. Nella prima versione del titolo che avevamo concepito volevamo sintetizzare l'intero spirito del libro. E questo perchè una esigenza che sentivamo come molto pressante era quella di circoscrivere da subito gli argomenti trattati e l'approccio seguito. La Statistica è infatti una disciplina assai ampia, nel senso che l'elenco dei concetti e dei metodi che le sono propri è molto vario e composito. In particolare è indubbia l'importanza della Statistica nel mondo della ricerca in Medicina, Biologia, Psicologia, Agricoltura, Sociologia, dove sia l'impostazione dell'esperimento che la stesura dei risultati hanno bisogno di un linguaggio *ad hoc*. Questo linguaggio è offerto proprio dalla Statistica inferenziale, che per noi assurge al ruolo di *modus operandi* della ricerca e di modalità di conoscenza dei fenomeni. Su questi temi torneremo diffusamente nel primo capitolo. Qui ci preme precisare che nel presente volume non ci occuperemo di tecniche di rilevazione campionaria. Pertanto non faremo alcun riferimento, se non sullo sfondo, in modo molto discreto e con la sola eccezione del campione *bernoulliano*, alla teoria del campionamento. Alle tecniche di indagine dedicheremo uno specifico volume, attualmente in preparazione, della giovane collana *Statistica con* R (quello che state sfogliando è il terzo dei volumi che la compongono). L'oggetto principale di questo manuale sono gli strumenti dell'inferenza statistica di maggiore utilizzo quando si devono confrontare due (o più di due) gruppi o campioni. Distingueremo fra campioni indipendenti e dipendenti così come chiariremo la differenza, che ricercheremo sostanzialmente nell'interpretazione dei risultati piuttosto che nei metodi utilizzati, tra gli studi osservazionali e quelli sperimentali.

In tutti i problemi che affronteremo il riferimento sarà una variabile di risposta, cioè la variabile di interesse oggetto dell'indagine. Lo scopo del ricercatore sarà dunque quello di confrontare i risultati ottenuti nei gruppi relativamente alla variabile di risposta. E tale confronto verrà instaurato sotto forma sia di un test di significatività sia di un intervallo di confidenza. Così in tutto il libro. Rispetto alla struttura di quest'ultimo, ci preme fare alcune brevi precisazioni. I primi quattro capitoli si occupano dei confronti cui abbiamo appena fatto cenno, mentre il quinto capitolo propone una riflessione generale su quanto esposto e presentato prima. Per la verità, il primo capitolo è propedeutico a quelli che seguono quanto a presentazione dei metodi utilizzati e della filosofia che intendiamo seguire. E in effetti vedremo che, trattando il caso di un solo campione, i confronti di cui parleremo non saranno instaurati fra due o più gruppi quanto piuttosto con una ipotesi di lavoro che il ricercatore vuole sottoporre a verifica.

Micciolo_2400pt.pdf 9 30/04/13 09:38

Comunque tutti i capitoli (a parte l'ultimo) sono strutturati allo stesso modo. Potete verificarlo immediatamente scorrendo il sommario. Questa struttura comune, che tra breve richiameremo per punti molto generali, fa di questo un libro, a nostro avviso, non standard. L'approccio che propone ci sembra, infatti, piuttosto originale e orientato a problemi dove le tecniche di inferenza statistica e la teoria che ne costituisce l'impalcatura di sostegno vengono introdotte dopo aver passato in rassegna una serie di esempi. Si tratta di casi di studio concreti, mutuati non solo da testi introduttivi o da problemi classici, ma anche dall'esperienza sul campo degli Autori e di alcuni loro Colleghi.

Gli esempi proposti a volte sono poco più di un pretesto per introdurre gli argomenti desiderati. Altre volte, però, detti esempi sono discussi anche nella interpretazione dei risultati cercando di far emergere eventuali limiti delle tecniche adottate.

Ci preme altresì ricordare che tutti gli esempi che prevedono, a vario livello, il disegno di simulazioni e l'esecuzione di esperimenti Monte Carlo sono perfettamente riproducibili nella forma proposta dal manuale: è sufficiente, a tale proposito, fissare il numero di innesco del generatore dei numeri pseudo-casuali utilizzando lo stesso "seme" che abbiamo utilizzato noi negli *script* presentati.

A questo proposito ricordiamo che la politica della collana *Statistica con* R è quella della massima diffusione del materiale prodotto. Pertanto, dal sito web di cui il volume si avvale (http://www.apogeonline.com/libri/9788850331949/scheda), sono scaricabili i *data-set* utilizzati, gli *script* R da noi proposti, i relativi *output* nonché dell'utile materiale integrativo. Un uso compiuto del presente manuale prevede di scaricare (dallo stesso sito web) e istallare un pacchetto R: la libreria `rmf`, la quale contiene la maggior parte delle funzioni R che utilizziamo nel volume. Si tratta di un pacchetto che abbiamo sviluppato principalmente per la didattica e che abbiamo già presentato nel primo volume della collana (*Analisi esplorativa dei dati con* R).

Dicevamo della struttura comune dei primi quattro capitoli. Ciascuno di questi si apre, come già accennato, con una disamina di esempi e con una analisi puntuale delle analogie che ce li fanno collocare nello stesso capitolo e delle differenze che ne impongono trattazione ed analisi separate. In generale, in ciascuno dei primi quattro capitoli si parla poi di medie, di metodi non parametrici e di metodi di ricampionamento. Solo a questo punto si affrontano le questioni relative alle proporzioni ed, infine, ai conteggi.

Ora dovrebbero cominciare a chiarirsi alcune delle ragioni per le quali parlavamo prima di "originalità" dell'approccio proposto. Innanzitutto la struttura imposta al volume privilegia decisamente, cosa per nulla frequente, il punto di vista del ricercatore. Infatti, con uno schema comune, differenziamo i vari capitoli sulla base del disegno dello studio: un campione, due campioni indipendenti, più di due campioni indipendenti e, infine, i campioni dipendenti. Per ogni disegno contemplato, lo ribadiamo, si passa poi alle analisi secondo lo schema già presentato e discusso.

Secondo aspetto che rende questo un manuale "diverso" e, a nostro avviso, originale. Come ormai risulta evidente, oltre ad argomenti standard di un libro di base

di inferenza statistica, abbiamo trattato anche metodi *computer-intensive* per sfruttare a pieno la potenza e la duttilità di R. Tra i metodi in parola spiccano quelli di ricampionamento, che tipicamente non vengono affrontati nei corsi istituzionali. Qualora, quindi, questo manuale venisse utilizzato in corsi di base attivati all'interno di lauree triennali, argomenti quali i metodi di ricampionamento (ma anche i metodi non parametrici) possono essere omessi senza in alcun modo snaturare la struttura modulare dei temi trattati nel libro. Proprio per la sua struttura, però, l'utilizzo che del manuale può essere fatto è duplice: una sua adozione più completa o con selezioni mirate sugli argomenti più complessi può essere concepita nelle lauree magistrali o nei *master* di primo o secondo livello.

L'approccio che abbiamo seguito nel libro ci offre il terzo spunto per definire questo un volume "diverso". La trattazione prevede ovunque di affrontare prima i test di verifica statistica delle ipotesi e, solo successivamente, di parlare delle stime intervallari. La consuetudine dei nostri tempi è quella di fare esattamente il contrario. I test di ipotesi e gli intervalli di confidenza per proporzioni e conteggi, poi, sono discussi molto in dettaglio, presentando, aspetto questo invero molto poco frequente nella manualistica di settore, i limiti dell'approssimazione gaussiana e alcune soluzioni (anche proposte molto di recente nella letteratura specialistica) che si possono trovare in specifiche librerie di R.

Dunque quello che state sfogliando è un libro non ovunque semplice e che prevede da parte del lettore almeno curiosità e spirito critico. A volte anche esperienza sul campo. Nonostante ciò la trattazione si mantiene (con la sola eccezione degli argomenti più complessi di cui abbiamo appena parlato) ad un livello affrontabile da lettori dotati delle conoscenze di base maturabili in un corso iniziale di Statistica. In particolare riteniamo che la conoscenza delle distribuzioni binomiale, di Poisson e normale siano, assieme al concetto di distribuzione campionaria, un bagaglio culturale imprescindibile per lo studio del presente manuale. Il nostro *Problemi ed esperimenti di statistica con* R costituisce al riguardo un utile riferimento.

Nel dare alle stampe il volume cogliamo l'occasione per ringraziare chi ha collaborato, in vario modo, alla sua realizzazione.

Innanzitutto dobbiamo ringraziare coloro che, con preziosi consigli e suggerimenti, ci hanno consentito di rendere questo manuale più efficace e, soprattutto, più conforme agli obiettivi che ci eravamo prefissati quando abbiamo cominciato a scriverlo: Maria Michela Dickson, Giorgio Fazio, Diego Giuliani, Davide Piacentino e Paolo Postiglione.

Gli studenti dei corsi di *Analisi dei dati e Statistica* e *Statistica, Probabilità e Inferenza* dell'ex Facoltà di Economia dell'Università di Trento (ora Dipartimento di Economia e Management), di *Statistica*, *Metodi quantitativi* e *Statistica per la ricerca sociale* dell'ex Facoltà di Sociologia dell'Università di Trento (ora Dipartimento di Sociologia e Ricerca Sociale), nonché i loro colleghi del corso di *Statistica per le Scienze Giuridiche* della ex Facoltà di Giurisprudenza (ora Dipartimento di Giurisprudenza) della LUISS "Guido Carli" di Roma (anni accademici 2010-2011, 2011-2012 e 2012-2013), vanno ringraziati per aver rilevato alcune imprecisioni contenute nelle

prime versioni e stesure distribuite loro in forma di dispensa. In particolare ci preme ricordare Valerio Pisoni, Massimo Iori, Claudio Fronterrè, Carlo Giovannini, Fabrizio Busnengo, Nicola Iavarone, Giulia Gaddari, Sara Cicatiello, Alberto Parrella, Sabato Vinci, Neli Marinova Nikolova e Maria Coclite.

Ringraziamo, infine, l'editore Apogeo e, in particolare, Alberto Kratter Thaler, Stefano Fabiano e Patrizia Villani, che continuano a credere nel nostro progetto e che continuano a seguirci con professionalità (e con molta pazienza) in tutte le fasi del nostro lavoro di autori e di curatori della collana *Statistica con* R. Silvia Maschio, della CompoMat s.r.l., va ringraziata per il suo insostituibile supporto nella predisposizione e nella gestione della classe LaTeXcon cui è stato scritto il volume. L'intera collana continua a fare riferimento allo stile editoriale concepito da Silvia, che ringraziamo per la sua infinita disponibilità e squisita gentilezza.

Questo libro è alla prima edizione e potrebbe contenere refusi, imperfezioni o errori sfuggiti al processo di revisione. Ci scusiamo in anticipo per questo e saremo grati a chi volesse segnalarceli all'indirizzo *e-mail* giuseppe.espa@unitn.it, in modo da predisporre un'eventuale *errata corrige* da depositare sul sito *web*.

Desideriamo ricordare con amore e gratitudine i nostri genitori: Ofelia e Pietro (Luisa Canal), Maria Luisa e Salvatore (Giuseppe Espa) e Celesta e Michelarcangelo (Rocco Micciolo). Questo libro è dedicato a loro.

Giuseppe Espa rivolge (rigorosamente in ordine alfabetico) un ringraziamento particolare a Maria Bellizzi, Cristiana Bergamini, Tatiana Boi, Stefano Calzolari, Adriano Decarli, Cesare Grandi, Lisa Licitra, Salvatore Mussari, Paolo Segatta e Franco Urbani per averlo accompagnato per mano fuori da un lungo e orribile tunnel. L'impagabile impegno e l'irrinunciabile sostegno di queste persone davvero speciali gli hanno consentito di ultimare il volume e chissà quanti altri progetti ancora.

Trento, 29 marzo 2013

Giuseppe Espa
Rocco Micciolo
Luisa Canal