Prefazione

Questo libro presenta una panoramica sui *fondamenti* – i più importanti concetti e risultati – della scienza statistica. Il pubblico a cui si rivolge è quello degli studenti che stanno seguendo un percorso per diventare data scientist. Tuttavia, questo non è un libro su come diventare un data scientist, o sui metodi più recenti usati dai data scientist, o su come analizzare i "big data" e l'ampia varietà di tipi di dati che i data scientist devono affrontare. È un libro il cui scopo è quello di insegnare ai potenziali data scientist i fondamenti di uno dei cardini della scienza dei dati (*data science*): la scienza statistica.

Oggi la scienza statistica è una disciplina molto ampia che include molte diverse specialità. Questo libro tratta in particolare gli argomenti che riteniamo debbano essere familiari a qualsiasi data scientist: metodi di statistica descrittiva, distribuzioni di probabilità, i metodi di statistica inferenziale come gli intervalli di confidenza e i test di significatività, la modellazione lineare e lineare generalizzata. Affiancato a studi di informatica e matematica, questo libro fornisce i fondamenti necessari per consentire un'introduzione alla scienza statistica attraverso lo studio di sezioni specializzate della statistica[1].

Il libro presuppone qualche conoscenza di analisi matematica, concentrandosi su *perché un'analisi statistica funziona* e su *come svolgerla*. I corsi universitari di statistica che richiedono nozioni di analisi matematica spesso sono chiamati *statistica matematica*, ma preferiamo evitare di utilizzare questo termine perché non vogliamo che gli studenti pensino che la scienza statistica sia un sottocampo della matematica o che siano indispensabili nozioni di matematica complessa per riuscire a capire e ad applicare la scienza statistica. In effetti utilizziamo soltanto gli strumenti di base dell'analisi matematica, quali il calcolo differenziale e integrale, e soltanto per alcuni argomenti. Rispetto ai libri tradizionali di statistica matematica, questo libro pone meno enfasi su teoria della probabilità, derivazione di distribuzioni di probabilità di trasformazioni di variabili aleatorie, teoria delle decisioni, formulazione e dimostrazione di teoremi. Introduce alcuni argomenti moderni che normalmente non appaiono nei testi tradizionali ma sono particolarmente importanti per i data scientist, come i modelli lineari generalizzati per risposte non normali, l'adattamento di modelli bayesiano e regolarizzato, la classificazione e il clustering. In ogni caso, la principale differenza tra questo libro e un testo tradizionale di statistica matematica è che qui mostriamo come implementare metodi statistici con software moderni e illustriamo concetti e teoria della statistica facendo ricorso a simulazioni.

Per usare e interpretare in modo appropriato i metodi della scienza statistica moderna, le competenze computazionali sono importanti quanto quelle matematiche. Oltre a utilizzare la matematica per mostrare "perché funziona", utilizziamo simulazioni computazionali e app disponibili su Internet per fornire spunti su risultati fondamentali quali il comportamento di distribuzioni campionarie e tassi di errore per inferenze statistiche.

[1] Come analisi multivariata, non parametrica, categoriale, progettazione e metodi di indagini campionarie, serie temporali, analisi di dati longitudinali, analisi di sopravvivenza, teoria delle decisioni, statistica bayesiana, modellazione stocastica, metodi computazionali della statistica, smoothing e modellazione non lineare.

In tutto il libro numerosi esempi con dati reali mostrano come utilizzare il software libero R per implementare metodi statistici. Inoltre, un'appendice presenta in maggiore dettaglio R e un'altra introduce l'uso di Python per eseguire analisi statistiche. L'appendice su Python illustra analisi per gli esempi trattati nei capitoli utilizzando R, perciò un docente può facilmente utilizzare il testo in un corso che utilizza Python come software principale. Poiché il libro è focalizzato sui fondamenti della scienza statistica, pone minore enfasi su alcuni aspetti pratici dell'analisi dei dati, come la preparazione e la pulizia dei file di dati. Tuttavia, le appendici dedicate ai software presentano anche analisi aggiuntive che fanno da supplemento agli esempi presentati nei capitoli. Un sito web dedicato al libro e aggiornato regolarmente, http://stat4ds.rwth-aachen.de, contiene tutti i file di dati analizzati e le appendici dedicate ai software in versioni più estese, oltre a un'altra appendice dedicata all'uso di Matlab per analisi statistiche. I file di dati sono disponibili anche su www.stat.ufl.edu/~aa/ds e presso il sito GitHub https://github.com/stat4DS/data.

Utilizzo come libro di testo per un corso

I Capitoli 1–6 di questo libro sono pensati come un libro di testo per un corso introduttivo alla scienza statistica dedicato a studenti universitari di corsi di laurea in scienza dei dati, statistica o matematica. A discrezione dei docenti è possibile saltare alcuni dei materiali meno centrali o più tecnici, come i Paragrafi 3.4, 4.9, 5.7, 5.8 e 6.7 (questi e altri paragrafi considerati facoltativi sono contrassegnati da un asterisco * accanto al titolo). Il complesso di tutti i nove capitoli e delle appendici su R e Python è appropriato anche per due corsi in sequenza. Il libro può essere utilizzato anche per corsi in cui la scienza statistica assume notevole rilievo, come l'econometria e la ricerca operativa. Potrebbe anche essere utile a studenti già laureati in scienze sociali, biologiche e ambientali che scelgono la statistica come area non obbligatoria, i quali potranno apprendere i fondamenti alla base dei metodi statistici che utilizzano. Un docente può utilizzare indifferentemente R o Python come software principale per il corso, dato che gli esempi presentati nei capitoli utilizzano R ma sono anche riportati con Python nell'appendice dedicata a tale linguaggio.

Ogni capitolo contiene molti esercizi con cui fare pratica e ampliare la teoria e i metodi. Gli esercizi sono raggruppati in due gruppi: quelli di *Analisi dei dati e applicazioni* richiedono che gli studenti eseguano analisi di dati simili a quelle presentate nel rispettivo capitolo, mentre quelli di *Metodi e concetti* attengono direttamente ai fondamenti trattati nel libro. Pongono domande su proprietà di metodi statistici, domande concettuali sulle loro basi e forniscono anche estensioni dei risultati riportati nel capitolo. Un'appendice riporta schemi di soluzioni per gli esercizi con numero dispari.

Questo libro non va inteso come una panoramica completa sulla scienza statistica: la disciplina è molto ampia e si estende ogni anno, anche con l'introduzione di nuovi campi in fase di sviluppo oggi e che non esistevano nemmeno nel ventesimo secolo. Tuttavia, riteniamo che fornisca una solida introduzione ai temi centrali che a nostro parere ogni data scientist dovrebbe conoscere.

Nella stesura del libro, Agresti (agresti@ufl.edu) ha assunto la responsabilità principale per i materiali dei capitoli e Kateri (maria.kateri@rwth-aachen.de) quella per le appendici su R e Python, anche nelle versioni più estese disponibili online, oltre all'appendice su Matlab presente sul sito del libro. Saremo grati per ogni commento o suggerimento che vogliate inviarci e di cui potremo tenere conto nelle edizioni future.

Ringraziamenti
Grazie ai numerosi amici e colleghi che hanno fornito commenti su varie versioni del testo, o data set o altri tipi di contributi, in particolare Alessandra Brazzale, Jane Brockmann, Brian Caffo, Sir David Cox, Bianca De Stavola, Cristina Cuesta, Travis Gerke, Sabrina Giordano, Anna Gottard, Ralitza Gueorguieva, Bernhard Klingenberg, Bhramar Mukherjee, Ranjini Natarajan, Madan Oli, Euijung Ryu, Alessandra Salvan, Nicola Sartori, Elena Stanghellini, Stephen Stigler, Gerhard Tutz, Roberta Varriale, Larry Winner e Daniela Witten. Grazie a Hassan Satvat per l'aiuto nell'impostazione del sito web del libro e a Bernhard Klingenberg per aver sviluppato le eccellenti app disponibili presso `www.artofstat.com` e spesso citate nel libro. Ringraziamo Joyce Robbins, Mintaek Lee, Jason M. Graham, Christopher Gaffney, Tumulesh Solanky e Steve Chung per le revisioni del nostro manoscritto eseguite per l'editore CRC Press. Infine, un ringraziamento speciale a John Kimmel, Executive Editor of Statistics per Chapman & Hall/CRC Press, per l'incoraggiamento e il sostegno a questo progetto.

Alan Agresti e Maria Kateri
Gainesville Florida e Brookline Massachusetts, USA;
Aachen, Germania

Aprile 2021