

# Sommario

<b>Presentazione dell'edizione italiana.....</b>	<b>XV</b>
<b>Prefazione.....</b>	<b>XVII</b>
<b>1 Introduzione alla scienza statistica.....</b>	<b>1</b>
1.1 Statistica: descrizioni e inferenze .....	1
1.1.1 Progettazione, statistica descrittiva e statistica inferenziale.....	2
1.1.2 Popolazioni e campioni .....	3
1.1.3 Parametri: sintesi numeriche della popolazione .....	3
1.1.4 Popolazione effettiva e popolazione concettuale.....	4
1.2 Tipi di dati e variabili.....	4
1.2.1 File di dati.....	4
1.2.2 Un esempio: General Social Survey (GSS) .....	5
1.2.3 Variabili .....	6
1.2.4 Variabili quantitative e variabili categoriche.....	7
1.2.5 Variabili discrete e variabili continue .....	7
1.2.6 Associazioni: variabili di risposta e variabili esplicative.....	8
1.3 Raccolta dei dati e casualizzazione.....	8
1.3.1 Casualizzazione.....	8
1.3.2 Raccogliere dati con un'indagine campionaria .....	9
1.3.3 Raccogliere dati con un esperimento.....	10
1.3.4 Raccogliere dati con uno studio osservazionale .....	10
1.3.5 Stabilire cause ed effetti: studi osservazionali e studi sperimentali.....	11
1.4 Statistiche descrittive: sintetizzare i dati.....	11
1.4.1 Esempio: emissioni di anidride carbonica nei paesi europei .....	12
1.4.2 Distribuzione di frequenza e istogramma.....	12
1.4.3 Descrivere il centro dei dati: media e mediana.....	13
1.4.4 Descrivere la variabilità dei dati: deviazione standard e varianza .....	14
1.4.5 Descrivere la posizione: percentili, quartili e box plot.....	16
1.5 Statistiche descrittive: sintetizzare dati multivariati.....	18
1.5.1 Dati quantitativi bivariati: grafico a dispersione, correlazione e regressione.....	18
1.5.2 Dati categorici bivariati: tabelle di contingenza .....	19
1.5.3 Statistiche descrittive per campioni e per popolazioni.....	20
1.6 Riepilogo.....	21

<b>2</b>	<b>Distribuzioni di probabilità .....</b>	<b>31</b>
2.1	Introduzione alla probabilità.....	31
2.1.1	Probabilità e frequenze relative in sequenze prolungate .....	31
2.1.2	Spazi campionari ed eventi .....	33
2.1.3	Assiomi della probabilità e conseguenti regole della probabilità .....	34
2.1.4	Esempio: diagnostica per lo screening di una malattia.....	35
2.1.5	Il teorema di Bayes .....	37
2.1.6	Legge moltiplicativa delle probabilità ed eventi indipendenti .....	37
2.2	Variabili aleatorie e distribuzioni di probabilità.....	38
2.2.1	Distribuzioni di probabilità per variabili aleatorie discrete.....	39
2.2.2	Esempio: distribuzione geometrica di probabilità .....	40
2.2.3	Distribuzioni di probabilità per variabili aleatorie continue .....	41
2.2.4	Esempio: distribuzione uniforme .....	41
2.2.5	Funzioni di probabilità ( <i>pdf</i> , <i>pmf</i> ) e funzione di distribuzione cumulata ( <i>cdf</i> ).....	42
2.2.6	Esempio: variabile aleatoria esponenziale.....	43
2.2.7	Famiglie di distribuzioni di probabilità con parametri indice.....	44
2.3	Valori attesi delle variabili aleatorie .....	44
2.3.1	Valore atteso e variabilità di una variabile aleatoria discreta.....	45
2.3.2	Valori attesi per le variabili aleatorie continue .....	46
2.3.3	Esempio: media e variabilità per una variabile aleatoria uniforme .....	47
2.3.4	Momenti superiori: asimmetria .....	48
2.3.5	Valori attesi di funzioni lineari di variabili aleatorie.....	48
2.3.6	Standardizzare una variabile aleatoria .....	49
2.4	Distribuzioni di probabilità discrete .....	49
2.4.1	Distribuzione binomiale .....	49
2.4.2	Esempio: composizione etnica di una giuria .....	50
2.4.3	Media, variabilità e asimmetria della distribuzione binomiale.....	51
2.4.4	Esempio: predire i risultati di un'indagine campionaria.....	53
2.4.5	La proporzione di campionamento come variabile aleatoria binomiale proporzionale.....	53
2.4.6	Distribuzione di Poisson .....	54
2.4.7	Variabilità e sovradisersione per la distribuzione di Poisson.....	55
2.5	Distribuzioni di probabilità continue.....	56
2.5.1	La distribuzione normale .....	56
2.5.2	La distribuzione normale standard.....	58
2.5.3	Esempi: trovare probabilità e percentili .....	58
2.5.4	La distribuzione gamma.....	59
2.5.5	La distribuzione esponenziale e i processi di Poisson .....	61
2.5.6	Quantili di una distribuzione di probabilità .....	62

2.5.7	Uso di una variabile aleatoria uniforme per generare casualmente una variabile aleatoria continua .....	63
2.6	Distribuzioni congiunte, condizionate e indipendenza .....	63
2.6.1	Distribuzioni di probabilità congiunte e marginali.....	63
2.6.2	Esempio: distribuzioni congiunte e marginali di felicità e reddito familiare.....	64
2.6.3	Distribuzioni di probabilità condizionate .....	64
2.6.4	Prove con categorie multiple: la distribuzione multinomiale .....	65
2.6.5	Valori attesi di somme di variabili aleatorie .....	67
2.6.6	Indipendenza di variabili aleatorie .....	67
2.6.7	Catena di Markov e indipendenza condizionata.....	68
2.7	Correlazione tra variabili aleatorie .....	69
2.7.1	Covarianza e correlazione.....	69
2.7.2	Esempio: correlazione tra reddito e felicità.....	70
2.7.3	L'indipendenza implica correlazione zero, ma non viceversa.....	71
2.7.4	Distribuzione normale bivariata .....	71
2.8	Riepilogo.....	73
<b>3</b>	<b>Distribuzioni campionarie.....</b>	<b>87</b>
3.1	Le distribuzioni campionarie: distribuzioni di probabilità delle statistiche .....	87
3.1.1	Esempio: predire i risultati di un'elezione con un exit poll .....	87
3.1.2	Distribuzione campionaria: variabilità di una statistica tra i campioni .....	89
3.1.3	Costruire una distribuzione campionaria.....	90
3.1.4	Esempio: simulazione per stimare il fatturato medio di un ristorante .....	91
3.2	Distribuzioni campionarie delle medie dei campioni.....	93
3.2.1	Media e varianza della media del campione di variabili casuali .....	93
3.2.2	Errore standard di una statistica.....	94
3.2.3	Esempio: errore standard del fatturato medio del campione.....	95
3.2.4	Esempio: errore standard della proporzione del campione in un exit poll.....	95
3.2.5	Legge dei grandi numeri: la media del campione converge alla media della popolazione .....	95
3.2.6	Le somme di variabili aleatorie normali, binomiali e di Poisson hanno la stessa distribuzione.....	96
3.3	Teorema del limite centrale: distribuzione campionaria normale per campioni di grandi dimensioni .....	97
3.3.1	La distribuzione campionaria della media del campione è approssimata da una distribuzione normale .....	97
3.3.2	Simulazioni per illustrare la distribuzione campionaria normale nel teorema del limite centrale .....	99
3.3.3	Riepilogo: popolazione, dati del campione e distribuzione campionaria .....	101
3.4	Distribuzioni campionarie normali per molte statistiche in grandi campioni .....	102

3.4.1	Il metodo delta.....	102
3.4.2	Il metodo delta applicato alla radice quadrata di Poisson per stabilizzare la varianza.....	103
3.4.3	Simulare distribuzioni campionarie di altre statistiche.....	104
3.4.4	Il ruolo chiave delle distribuzioni campionarie nell'inferenza statistica.....	106
3.5	Riepilogo.....	106
<b>4</b>	<b>Inferenza statistica: stima .....</b>	<b>115</b>
4.1	Stime puntuali e intervalli di confidenza.....	115
4.1.1	Proprietà degli stimatori: non distorsione, consistenza, efficienza.....	116
4.1.2	Valutare le proprietà degli stimatori.....	117
4.1.3	Stima intervallare: intervalli di confidenza per parametri.....	118
4.2	Funzione di verosimiglianza e stima di massima verosimiglianza.....	118
4.2.1	Funzione di verosimiglianza.....	118
4.2.2	Metodo di stima di massima verosimiglianza.....	120
4.2.3	Proprietà degli stimatori di massima verosimiglianza (ML).....	121
4.2.4	Esempio: varianza dello stimatore ML di un parametro binomiale.....	122
4.2.5	Esempio: varianza dello stimatore ML di una media di Poisson.....	122
4.2.6	Sufficienza e invarianza per stime ML.....	123
4.3	Costruzione di intervalli di confidenza.....	124
4.3.1	Uso di una quantità pivotale per indurre un intervallo di confidenza.....	124
4.3.2	Un intervallo di confidenza in grandi campioni per la media.....	126
4.3.3	Intervalli di confidenza per proporzioni.....	126
4.3.4	Esempio: atei e agnostici in Europa.....	128
4.3.5	Uso della simulazione per illustrare le prestazioni degli intervalli di confidenza nel lungo termine.....	128
4.3.6	Determinare la dimensione del campione prima di raccogliere i dati.....	129
4.3.7	Esempio: dimensione del campione per valutare una strategia pubblicitaria.....	130
4.4	Intervalli di confidenza per medie di popolazioni normali.....	131
4.4.1	La distribuzione $t$ .....	131
4.4.2	Intervallo di confidenza per una media con la distribuzione $t$ .....	133
4.4.3	Esempio: stima della variazione di peso media per ragazze anoressiche.....	133
4.4.4	Robustezza a violazioni dell'assunzione di popolazione normale.....	134
4.4.5	Costruzione di una distribuzione $t$ usando distribuzioni chi-quadro e normale standard.....	135
4.4.6	Perché la quantità pivotale ha la distribuzione $t$ ?.....	137
4.4.7	Distribuzione di Cauchy: distribuzione $t$ con $df = 1$ di comportamento anomalo.....	138
4.5	Confronto tra due medie o proporzioni della popolazione.....	138
4.5.1	Un modello per confrontare medie: normalità con variabilità comune.....	139
4.5.2	Errore standard e intervallo di confidenza per confronto tra medie.....	139

4.5.3	Esempio: confrontare un gruppo di trattamento e uno di controllo .....	140
4.5.4	Intervallo di confidenza per confrontare due proporzioni .....	142
4.5.5	Esempio: le preghiere aiutano i pazienti di chirurgia coronarica? .....	143
4.6	Bootstrap .....	144
4.6.1	Ricampionamento computazionale e intervalli di confidenza bootstrap .....	144
4.6.2	Esempio: intervalli di confidenza bootstrap per dati di biblioteche .....	145
4.7	L'approccio bayesiano all'inferenza statistica .....	147
4.7.1	Distribuzioni a priori e a posteriori bayesiane .....	148
4.7.2	Inferenza binomiale bayesiana: distribuzioni a priori beta .....	149
4.7.3	Esempio: credere nell'inferno.....	150
4.7.4	Interpretazione: intervalli bayesiani e classici a confronto .....	150
4.7.5	Intervallo a posteriori bayesiano per il confronto di proporzioni .....	151
4.7.6	Intervalli a posteriori HPD (Highest Posterior Density).....	151
4.8	Inferenza bayesiana per medie .....	152
4.8.1	Inferenza bayesiana per una media normale .....	152
4.8.2	Esempio: analisi bayesiana per la terapia dell'anoressia .....	153
4.8.3	Inferenza bayesiana per medie normali con distribuzioni a priori improprie .....	154
4.8.4	Predire un'osservazione futura: distribuzione predittiva bayesiana.....	155
4.8.5	Prospettiva bayesiana, approccio bayesiano empirico e approccio bayesiano gerarchico .....	155
4.9	Motivi alla base delle buone prestazioni degli stimatori di massima verosimiglianza e di Bayes .....	156
4.9.1	Gli stimatori ML hanno distribuzioni normali in grandi campioni .....	156
4.9.2	L'efficienza asintotica di stimatori ML è pari a quella dei migliori stimatori non distorti .....	159
4.9.3	Gli stimatori di Bayes hanno buone prestazioni anche in grandi campioni.....	160
4.9.4	Il principio di verosimiglianza.....	160
4.10	Riepilogo.....	160
<b>5</b>	<b>Inferenza statistica: test di significatività.....</b>	<b>177</b>
5.1	Elementi di un test di significatività .....	177
5.1.1	Esempio: test per distorsione nella selezione di manager.....	177
5.1.2	Assunzioni, ipotesi, statistica test, <i>p-value</i> e conclusione.....	178
5.2	Test di significatività per proporzioni e medie.....	181
5.2.1	Elementi di un test di significatività per una proporzione.....	181
5.2.2	Esempio: il cambiamento climatico è una grave minaccia? .....	182
5.2.3	Test di significatività unilaterali .....	183
5.2.4	Elementi di un test di significatività per una media .....	184
5.2.5	Esempio: test di significatività sull'orientamento politico .....	185

5.3	Test di significatività per il confronto tra medie .....	187
5.3.1	Test di significatività per la differenza tra due medie.....	187
5.3.2	Esempio: confronto tra gruppo di trattamento e gruppo di controllo .....	188
5.3.3	Ammontare dell'effetto per il confronto tra due medie.....	189
5.3.4	Inferenza bayesiana per il confronto tra due medie.....	190
5.3.5	Esempio: confronto bayesiano tra gruppo di trattamento e di controllo .....	191
5.4	Test di significatività per il confronto tra proporzioni .....	191
5.4.1	Test di significatività per la differenza tra due proporzioni.....	192
5.4.2	Esempio: confronto tra pazienti chirurgici con preghiere e senza.....	192
5.4.3	Inferenza bayesiana per il confronto di due proporzioni.....	193
5.4.4	Test chi quadro per proporzioni multiple in tabelle di contingenza .....	194
5.4.5	Esempio: felicità e stato civile.....	195
5.4.6	Residui standardizzati: descrivere la natura di un'associazione.....	196
5.5	Test di significatività: decisioni ed errori.....	198
5.5.1	Il livello alfa: prendere una decisione in base al <i>p-value</i> .....	198
5.5.2	Mai "Accettare $H_0$ " in un test di significatività .....	199
5.5.3	Errori di Tipo I e di Tipo II .....	199
5.5.4	Al diminuire di $P(\text{Errore di Tipo I})$ , aumenta $P(\text{Errore di Tipo II})$ .....	200
5.5.5	Esempio: test su elementi di verità dell'astrologia.....	201
5.5.6	Potenza di un test.....	203
5.5.7	Prendere decisioni o riportare il <i>p-value</i> .....	204
5.6	Dualità fra test di significatività e intervalli di confidenza.....	204
5.6.1	Connessione fra test bilaterali e intervalli di confidenza .....	204
5.6.2	Effetto della dimensione del campione: significatività statistica e pratica .....	206
5.6.3	I test di significatività sono meno utili degli intervalli di confidenza.....	207
5.6.4	Test di significatività e <i>p-value</i> possono essere fuorvianti .....	207
5.7	Test del rapporto di verosimiglianza e intervalli di confidenza .....	209
5.7.1	Rapporto di verosimiglianza e test chi quadro.....	209
5.7.2	Test del rapporto di verosimiglianza e intervallo di confidenza per una proporzione.....	210
5.7.3	Triade di test: del rapporto di verosimiglianza, di Wald, <i>score test</i> .....	211
5.8	Test non parametrici .....	213
5.8.1	Un test di permutazione per confrontare due gruppi.....	213
5.8.2	Esempio: carezze vs parole dolci per i cani.....	213
5.8.3	Test di Wilcoxon: confrontare ranghi di medie per due gruppi.....	215
5.8.4	Confrontare distribuzioni del tempo di sopravvivenza con dati censurati .....	216
5.9	Riepilogo.....	219

<b>6</b>	<b>Modelli lineari e minimi quadrati .....</b>	<b>233</b>
6.1	Modello di regressione lineare e fit dei minimi quadrati .....	233
6.1.1	Il modello lineare descrive un'aspettativa condizionata .....	233
6.1.2	Descrivere la variazione attorno all'aspettativa condizionata .....	234
6.1.3	Adattamento (fitting) con il modello dei minimi quadrati .....	235
6.1.4	Esempio: modello lineare per le corse in montagna in Scozia.....	237
6.1.5	Correlazione .....	238
6.1.6	Regressione verso la media nei modelli di regressione lineare.....	239
6.1.7	Modelli lineari e realtà .....	240
6.2	Regressione multipla: modelli lineari con variabili esplicative multiple.....	241
6.2.1	Interpretazione di effetti nei modelli di regressione multipla .....	241
6.2.2	Esempio: regressione multipla per le corse in montagna in Scozia .....	242
6.2.3	Associazione e causazione .....	242
6.2.4	Confondimento, associazione spuria e indipendenza condizionata .....	244
6.2.5	Esempio: modellazione del tasso di criminalità in Florida.....	245
6.2.6	Equazioni per le stime di minimi quadrati nella regressione multipla .....	246
6.2.7	Interazione tra variabili esplicative nei loro effetti.....	247
6.2.8	Distanza di Cook: rilevare osservazioni insolite e influenti.....	249
6.3	Riepilogo sulla variabilità nei modelli di regressione lineare.....	250
6.3.1	Varianza di errore e chi-quadro per modelli lineari .....	251
6.3.2	Scomporre la variabilità in modello spiegato e parti non spiegate.....	251
6.3.3	$R$ quadro e correlazione multipla.....	253
6.3.4	Esempio: $R$ quadro per modellare le corse in montagna in Scozia.....	253
6.4	Inferenza statistica per modelli lineari normali.....	254
6.4.1	La distribuzione $F$ : test che tutti gli effetti sono uguali a 0 .....	255
6.4.2	Esempio: modello lineare normale per disturbo mentale.....	256
6.4.3	Test $t$ e intervalli di confidenza per effetti individuali .....	257
6.4.4	Multicollinearità: variabili esplicative quasi ridondanti.....	258
6.4.5	Intervallo di confidenza per $E(Y)$ e intervallo di previsione per $Y$ .....	259
6.4.6	Il test $F$ che tutti gli effetti sono uguali a zero è un test del rapporto di verosimiglianza .....	261
6.5	Variabili esplicative categoriche in modelli lineari .....	262
6.5.1	Variabili indicatrici per categorie.....	262
6.5.2	Esempio: confrontare i redditi medi di gruppi etnici.....	263
6.5.3	Analisi della varianza (ANOVA): un test $F$ per confrontare più medie .....	264
6.5.4	Confronti multipli tra medie: metodi di Bonferroni e di Tukey .....	266
6.5.5	Modelli con variabili esplicative categoriche e quantitative .....	267
6.5.6	Confronto di due modelli lineari normali annidati .....	268
6.5.7	Interazione con variabili esplicative categoriche e quantitative .....	270

6.6	Inferenza bayesiana per modelli lineari normali .....	270
6.6.1	Distribuzioni a priori e a posteriori per modelli lineari normali.....	271
6.6.2	Esempio: modello lineare bayesiano per disturbo mentale.....	271
6.6.3	Approccio bayesiano al layout unidirezionale normale .....	272
6.7	Formulazione matriciale di modelli lineari .....	272
6.7.1	La matrice del modello .....	273
6.7.2	Stime dei minimi quadrati ed errori standard .....	273
6.7.3	Matrice con il cappello e leverage.....	274
6.7.4	Alternative ai minimi quadrati: regressione robusta e regolarizzazione .....	275
6.7.5	Ottimalità ristretta dei minimi quadrati: teorema di Gauss-Markov .....	275
6.7.6	Formulazione matriciale di un modello lineare normale bayesiano .....	276
6.8	Riepilogo.....	277
<b>7</b>	<b>Modelli lineari generalizzati .....</b>	<b>289</b>
7.1	Introduzione ai modelli lineari generalizzati.....	289
7.1.1	Le tre componenti di un modello lineare generalizzato.....	289
7.1.2	Modelli lineari generalizzati per distribuzione normale, binomiale e di Poisson .....	290
7.1.3	Esempio: modelli lineari generalizzati per i prezzi di vendita delle case.....	291
7.1.4	Devianza.....	293
7.1.5	Confronto del modello del rapporto di verosimiglianza con differenza di devianza.....	294
7.1.6	Selezione del modello: AIC e tradeoff distorsione/varianza .....	295
7.1.7	Vantaggi dei modelli lineari generalizzati rispetto alla trasformazione dei dati.....	297
7.1.8	Esempio: modelli lineari generalizzati normale e gamma per dati Covid-19 .....	298
7.2	Modello di regressione logistica per dati binari .....	299
7.2.1	Regressione logistica: espressioni del modello.....	299
7.2.2	Interpretazione dei parametri: effetti su probabilità e odds .....	300
7.2.3	Esempio: studio dose-risposta per tarme della farina.....	301
7.2.4	Dati binari raggruppati e non: effetti su stime e devianza.....	303
7.2.5	Esempio: modellare l'occupazione italiana con link logit e link identità.....	305
7.2.6	Separazione completa e stime logistiche infinite del parametro.....	307
7.3	Inferenza bayesiana per modelli lineari generalizzati .....	308
7.3.1	Distribuzioni a priori normali per parametri di modelli lineari generalizzati .....	309
7.3.2	Esempio: regressione logistica per affetti da cancro dell'endometrio.....	309
7.4	Modelli loglineari di Poisson per dati di conteggio.....	312
7.4.1	Modelli loglineari di Poisson .....	312
7.4.2	Esempio: modellare il numero di satelliti nei limuli.....	312
7.4.3	Modellazione di tassi: inserimento di un offset nel modello.....	314
7.4.4	Esempio: sopravvivenza al cancro polmonare .....	314



7.5	Modelli binomiali negativi per dati di conteggio sovradispersi.....	316
7.5.1	Varianza aumentata a causa dell'eterogeneità .....	316
7.5.2	Binomiale negativo: mistura gamma di distribuzioni di Poisson.....	317
7.5.3	Esempio: modello binomiale negativo per i dati sui limuli.....	318
7.6	Adattamento iterativo del modello GLM.....	319
7.6.1	Metodo di Newton–Raphson .....	319
7.6.2	Adattamento di Newton–Raphson di un modello di regressione logistica.....	321
7.6.3	Matrice di covarianza di stimatori dei parametri e Fisher scoring.....	322
7.6.4	Equazioni di verosimiglianza e matrice di covarianza per modelli lineari generalizzati di Poisson .....	323
7.7	Regolarizzazione con alto numero di parametri.....	324
7.7.1	Metodi di verosimiglianza penalizzata .....	324
7.7.2	Metodi di verosimiglianza penalizzata: lasso .....	325
7.7.3	Esempio: predire le opinioni con dati di un sondaggio tra studenti.....	326
7.7.4	Perché ridurre a 0 le stime ML?.....	327
7.7.5	Riduzione delle dimensioni: analisi delle componenti principali.....	328
7.7.6	Inferenza bayesiana con un grande numero di parametri .....	329
7.7.7	$n$ molto grande: gestire i big data .....	329
7.8	Riepilogo.....	330
<b>8</b>	<b>Classificazione e clustering.....</b>	<b>343</b>
8.1	Classificazione: analisi discriminante lineare e alberi.....	343
8.1.1	Classificazione con funzione discriminante lineare di Fisher .....	344
8.1.2	Esempio: predire se i limuli hanno satelliti.....	344
8.1.3	Riepilogo del potere predittivo: tabelle di classificazione e curve ROC .....	346
8.1.4	Alberi di classificazione: previsione grafica.....	348
8.1.5	Confronto tra regressione logistica, analisi discriminante lineare e alberi di classificazione .....	350
8.1.6	Altri metodi per la classificazione: $k$ vicini più prossimi e reti neurali.....	351
8.2	Analisi dei cluster .....	355
8.2.1	Misurazione di dissimilarità tra osservazioni su risposte binarie .....	355
8.2.2	Algoritmo di clustering gerarchico e suo dendrogramma .....	356
8.2.3	Esempio: clustering di stati USA sui risultati delle elezioni presidenziali .....	356
8.3	Riepilogo.....	359
<b>9</b>	<b>Storia della scienza statistica.....</b>	<b>365</b>
9.1	L'evoluzione della scienza statistica .....	365
9.1.1	Evoluzione della probabilità.....	365
9.1.2	Evoluzione della statistica descrittiva e inferenziale.....	366

9.2	I pilastri della conoscenza e della pratica statistica .....	369
9.2.1	I sette pilastri del sapere statistico di Stigler.....	369
9.2.2	I sette pilastri di conoscenza per praticare la scienza dei dati .....	371
<b>A</b>	<b>Uso di R nella scienza statistica .....</b>	<b>375</b>
A.0	Le basi di R.....	375
A.0.1	Avviare una sessione, inserire comandi e uscire.....	375
A.0.2	Installare e caricare i package di R.....	375
A.0.3	Funzioni e strutture dati in R .....	376
A.0.4	Input di dati in R.....	378
A.0.5	Controllo di flusso in R.....	379
A.1	Capitolo 1: R per la statistica descrittiva.....	379
A.1.1	Gestione dei dati e <i>wrangling</i> .....	379
A.1.2	Istogrammi e altri tipi di grafici.....	380
A.1.3	Statistica descrittiva .....	382
A.1.4	Valori mancanti nei file di dati .....	385
A.1.5	Sintesi di dati quantitativi bivariati.....	386
A.1.6	Sintesi di dati categorici bivariati .....	387
A.2	Capitolo 2: R per distribuzioni di probabilità.....	388
A.2.1	Funzioni di R per distribuzioni di probabilità .....	388
A.2.2	Quantili, grafici <i>Q-Q</i> e grafico normale quantile.....	389
A.2.3	Distribuzioni di probabilità congiunte e condizionate.....	392
A.3	Capitolo 3: R per distribuzioni campionarie.....	392
A.3.1	Simulare la distribuzione campionaria di una statistica.....	393
A.3.2	Simulazione Monte Carlo .....	394
A.4	Capitolo 4: R per le stime .....	396
A.4.1	Intervalli di confidenza per proporzioni.....	396
A.4.2	Intervalli di confidenza per medie di sottogruppi e differenze a coppie .....	396
A.4.3	La distribuzione <i>t</i> e altre distribuzioni di probabilità per l'inferenza statistica .....	397
A.4.4	Funzione di distribuzione cumulata empirica.....	397
A.4.5	Bootstrap non parametrico e parametrico .....	399
A.4.6	Intervalli HPD bayesiani per confrontare proporzioni .....	401
A.5	Capitolo 5: R per test di significatività.....	402
A.5.1	Fattori di Bayes e un test <i>t</i> di Bayes.....	402
A.5.2	Simulare la distribuzione esatta della statistica del rapporto di verosimiglianza .....	403
A.5.3	Statistiche non parametriche: test di permutazione e test di Wilcoxon.....	404
A.6	Capitolo 6: R per modelli lineari.....	405
A.6.1	Modelli lineari con la funzione <i>lm</i> .....	405
A.6.2	Grafici diagnostici per modelli lineari.....	405
A.6.3	Grafici per bande di regressione e distribuzioni a posteriori.....	407

A.7	Capitolo 7: <i>R</i> per modelli lineari generalizzati .....	408
A.7.1	La funzione <i>glm</i> .....	408
A.7.2	Tracciare un adattamento del modello di regressione logistica.....	408
A.7.3	Selezione del modello per modelli lineari generalizzati .....	408
A.7.4	Risposte correlate: modelli marginali, a effetti aleatori, transizionali.....	411
A.7.5	Modellazione di serie temporali .....	412
A.8	Capitolo 8: <i>R</i> per classificazione e clustering.....	414
A.8.1	Visualizzazione dei risultati dell'analisi discriminante lineare.....	414
A.8.2	Convalida incrociata e addestramento del modello .....	415
A.8.3	Alberi di classificazione e di regressione.....	416
A.8.4	Analisi dei cluster con variabili quantitative.....	417
<b>B</b>	<b>Usa di Python nella scienza statistica.....</b>	<b>419</b>
B.0	Le basi di Python .....	419
B.0.1	Nozioni preliminari su Python .....	419
B.0.2	Strutture dati e input di dati .....	420
B.1	Capitolo 1: Python per statistica descrittiva .....	421
B.1.1	Generazione di numeri casuali .....	421
B.1.2	Statistiche di riepilogo e grafici per variabili quantitative .....	421
B.1.3	Statistiche descrittive per dati quantitativi bivariati.....	422
B.1.4	Statistiche descrittive per dati categoriali bivariati .....	423
B.1.5	Simulazione di campioni tratti da una popolazione con distribuzione a campana ...	424
B.2	Capitolo 2: Python per distribuzioni di probabilità.....	425
B.2.1	Simulazione di una probabilità come frequenza relativa di lungo termine.....	425
B.2.2	Funzioni Python per distribuzioni di probabilità discrete.....	425
B.2.3	Funzioni Python per distribuzioni di probabilità continue .....	427
B.2.4	Valori attesi di variabili aleatorie .....	429
B.3	Capitolo 3: Python per distribuzioni campionarie .....	431
B.3.1	Simulazione per illustrare una distribuzione campionaria .....	431
B.3.2	Legge dei grandi numeri.....	431
B.4	Capitolo 4: Python per stime .....	431
B.4.1	Intervalli di confidenza per proporzioni.....	431
B.4.2	La distribuzione <i>t</i> .....	432
B.4.3	Intervalli di confidenza per medie.....	432
B.4.4	Intervalli di confidenza per confronto di medie e di proporzioni .....	433
B.4.5	Intervalli di confidenza bootstrap .....	434
B.4.6	Intervalli a posteriori bayesiani per proporzioni e medie .....	434
B.5	Capitolo 5: Python per test di significatività .....	435
B.5.1	Test di significatività per proporzioni .....	435

B.5.2	Test chi quadro per confrontare più proporzioni in tabelle di contingenza .....	436
B.5.3	Test di significatività per medie .....	436
B.5.4	Test di significatività per il confronto di medie .....	437
B.5.5	Potenza di un test di significatività .....	438
B.5.6	Statistiche non parametriche: test di permutazione e test di Wilcoxon .....	439
B.5.7	Stima di Kaplan-Meier di funzioni di sopravvivenza .....	439
B.6	Capitolo 6: Python per modelli lineari .....	440
B.6.1	Adattamento di modelli lineari .....	440
B.6.2	Correlazione e $R$ quadro .....	442
B.6.3	Diagnostica: residui e distanze di Cook per modelli lineari .....	442
B.6.4	Inferenza statistica e previsione per modelli lineari .....	446
B.6.5	Variabili esplicative categoriche in modelli lineari .....	447
B.6.6	Adattamento bayesiano di modelli lineari .....	448
B.7	Capitolo 7: Python per modelli lineari generalizzati .....	448
B.7.1	GLM con funzione link identità .....	449
B.7.2	Regressione logistica: link logit con dati binari .....	451
B.7.3	Separazione e adattamento bayesiano in regressione logistica .....	452
B.7.4	Modello loglineare di Poisson per conteggi .....	453
B.7.5	Modello binomiale negativo per dati di conteggio .....	455
B.7.6	Regularizzazione: regressione logistica penalizzata con il lasso .....	456
B.8	Capitolo 8: Python per classificazione e clustering .....	457
B.8.1	Analisi discriminante lineare .....	457
B.8.2	Alberi di classificazione e reti neurali per previsione .....	459
B.8.3	Analisi dei cluster .....	461
<b>C</b>	<b>Soluzioni sintetiche degli esercizi .....</b>	<b>463</b>
C.1	Capitolo 1: soluzioni degli esercizi .....	463
C.2	Capitolo 2: Soluzioni degli esercizi .....	465
C.3	Capitolo 3: soluzioni degli esercizi .....	468
C.4	Capitolo 4: soluzioni degli esercizi .....	470
C.5	Capitolo 5: soluzioni degli esercizi .....	474
C.6	Capitolo 6: soluzioni degli esercizi .....	478
C.7	Capitolo 7: soluzioni degli esercizi .....	482
C.8	Capitolo 8: soluzioni degli esercizi .....	485
<b>D</b>	<b>Bibliografia .....</b>	<b>487</b>

## Presentazione dell’edizione italiana

Di Fabio Corradi\*

*Statistica per data scientist* di Alan Agresti e Maria Kateri è, a mio parere, un testo che mancava nella vasta offerta di libri per l’università. Ciò che lo distingue, come testo per un primo corso di statistica, è l’indice, che comprende argomenti standard ma anche argomenti mai presentati a questo livello, quali un bellissimo capitolo sui modelli lineari generalizzati e un’introduzione alla classificazione seguendo sia l’approccio modellistico sia quello algoritmico. Di più: quasi tutti i temi vengono trattati sia con l’approccio frequentista sia con quello Bayesiano. L’approccio frequentista è illustrato in modo onesto e approfondito. L’approccio Bayesiano è ben illustrato ma forse poco valorizzato nei commenti dei risultati. Il tutto lascia al docente l’opportunità di operare i dovuti confronti fra le due scuole.

L’approccio didattico è quello di un testo facile con molti esercizi, alcuni fra i quali, il docente li riconoscerà, necessitano di un po’ più di teoria di quella illustrata nel testo. Insomma, con un’opportuna scelta nelle assegnazioni si può consolidare la materia esposta a lezione e lasciare gli studenti cimentarsi con qualcosa che, in fase di revisione, farà emergere altri preziosi frammenti di teoria.

Altre piccole perle: diversi accenni all’inferenza causale, e alla selezione di variabili fatta in modo «moderno» ovvero attraverso penalizzazioni. Peccato che in questo caso ci sia solo la versione frequentista e manchi quella Bayesiana delle a priori «spike and slab» che sarebbero ben servite per un ulteriore confronto che potrà sempre essere fatto dal docente integrando un po’ il testo. Infine ci sono due Appendici che illustrano l’uso di R e Python per la statistica. Le due appendici possono essere utilizzate in alternativa, ovvero scegliendo quella più adatta al background degli studenti e ai corsi a cui sono già stati esposti. Nel caso di un/una aspirante data scientist che avesse poca dimestichezza con la programmazione, questo è un viatico ideale perché si procede per esempi articolati sui capitoli del libro, seguendo l’approccio della mano guidata per scrivere le lettere dell’alfabeto.

La platea di studenti a cui il testo è rivolto è quella di corsi di laurea che abbiano avuto possibilmente un primo contatto con il calcolo delle probabilità. L’intero materiale presentato è idoneo per un corso di 9 crediti, ovvero una settantina di ore, essendo sempre possibili semplificazioni che lo riducano a un corso di circa cinquanta ore.

FABIO CORRADI

---

\*Fabio Corradi è professore ordinario di Statistica all’università di Firenze.