

Περιεχόμενα

Πρόλογος	xi
Ευχαριστίες	xiii

• Μέρος Πρώτο •

Κεφάλαιο 1	3
1.1 Σε ποιον απευθύνεται αυτό το βιβλίο	3
1.2 Στόχοι του βιβλίου.....	4
1.3 Περιεχόμενα.....	5
1.3.1 Προγραμματισμός στην R.....	6
1.3.2 Βασικές αρχές στατιστικής	6
1.3.3 Μεθοδολογίες ανάλυσης δεδομένων	6
1.3.4 Εφαρμογές σε πραγματικά δεδομένα.....	6
1.4 Πλεονεκτήματα της R	7
1.5 Ξεκινώντας με την R	8
1.5.1 Εγκαθιστώντας την R	8
1.5.2 Εξοικείωση με το περιβάλλον της R	8
1.6 Το ολοκληρωμένο περιβάλλον R-Studio	12
1.7 Πως είναι γραμμένο αυτό το βιβλίο.....	12
1.8 Διαβάστε περισσότερα	14
1.8.1 Βιβλία για την R.....	14
1.8.2 Διαδικτυακές πηγές.....	15
Κεφάλαιο 2 - Μεταβλητές, τελεστές και δεδομένα.....	17
2.1 Μεταβλητές στην R	17
2.1.1 Απλές μεταβλητές	17
2.1.2 Εκχώρηση τιμής σε μεταβλητή.....	18
2.1.3 Κανόνες ονομασίας μεταβλητών	19
2.1.4 Είδη μεταβλητών στην R	19

2.2 Τελεστές και απλές πράξεις με μεταβλητές	21
2.2.1 Αριθμητικές πράξεις	21
2.2.2 Λογικές πράξεις	23
2.3 Συναρτήσεις	24
2.4 Πολυδιάστατα δεδομένα	25
2.4.1 Έτοιμα (built-in) δεδομένα	25
2.5 Εισαγωγή δεδομένων από αρχεία	29
2.5.1 Ανάγνωση δεδομένων σε πίνακα	30
2.5.2 Ανάγνωση αρχείων κειμένου	31
2.5.3 Ορισμός φακέλου εργασίας και πλήρους μονοπατιού	31
2.6 Εγγραφή αποτελεσμάτων σε αρχεία	32
Συμπεράσματα	34

Κεφάλαιο 3 - Τύποι δεδομένων..... 35

Εισαγωγή	35
Τύποι δεδομένων στην R	35
3.1 Διανύσματα (vectors)	36
3.1.1 Γενικά χαρακτηριστικά διανυσμάτων	36
3.1.2 Δημιουργία διανυσμάτων και “εξαναγκασμός” (coersion)	38
3.1.3 Εξωτερικός “εξαναγκασμός” από τον χρήστη.	39
3.1.4 Συναρτήσεις διανυσμάτων	41
3.1.4.1 Συναρτήσεις αριθμητικών διανυσμάτων	41
3.1.4.2 Συναρτήσεις αλφαριθμητικών διανυσμάτων	44
3.2 Πράξεις σε διανύσματα	46
3.3 Πίνακες δύο διαστάσεων	48
3.3.1 Πρόσβαση σε στοιχεία πινάκων	50
3.3.2 Συναρτήσεις πινάκων	51
3.4 Πλαίσια δεδομένων	54
3.4.1 Εφαρμογή συναρτήσεων σε πίνακες και πλαίσια δεδομένων (apply)	55
3.5 Πίνακες περισσότερων από δύο διαστάσεων	57
3.6 Λίστες	59
3.6.1 Δημιουργία λιστών	59
3.6.2 Χειρισμός λιστών με συναρτήσεις τύπου <i>apply()</i>	62
Συμπεράσματα	63

Κεφάλαιο 4 - Πλαίσια δεδομένων, παράγοντες και χειρισμοί συνόλων.....	65
Εισαγωγή.....	65
4.1 Χειρισμοί πλαισίων δεδομένων	65
4.1.1 Κενές τιμές (missing values)	66
4.1.2 Συναρτήσεις ελέγχου τιμών	68
4.2 Παράγοντες (factors).....	73
4.2.1 Δημιουργία και χειρισμός επιπέδων (levels) σε παράγοντες.....	73
4.2.2 Χρήση παραγόντων σε πλαίσια δεδομένων	74
4.3 Δημιουργία υποσυνόλων (subsetting)	77
4.3.1 Δημιουργία υποσυνόλων με απλούς ελέγχους.....	78
4.3.2 Υποσύνολα με συνδυασμούς ελέγχων.....	78
4.3.3 Λήψη υποσυνόλων από πλαίσια δεδομένων με τη συνάρτηση <i>subset()</i>	79
4.3.4 Λήψη θέσεων στοιχείων υποσυνόλων με τη συνάρτηση <i>which()</i>	81
4.4 Πράξεις σε σύνολα	83
4.4.1 Ύπαρξη στοιχείων συνόλου σε άλλο σύνολο	85
4.4.2 Συγκρίσεις μεταξύ συνόλων σε επίπεδο θέσης.....	86
 Κεφάλαιο 5 - Γραφικές παραστάσεις	89
Εισαγωγή.....	89
5.1 Ραβδογράμματα με την <i>barplot()</i>	89
5.2 Προχωρημένα ραβδογράμματα.....	95
5.3 Πίτες με την <i>pie()</i>	102
5.4 Διαγράμματα σκέδασης για συζευγμένα δεδομένα με την <i>plot()</i>	104
5.5 Χρήση της <i>plot()</i> για σειριακά δεδομένα.	107
5.6 Πολλαπλά διαγράμματα σε μια γραφική με την <i>lines()</i>	110
5.7 Τρισδιάστατα δεδομένα με τις <i>persp()</i> και <i>filled.contour()</i>	113
5.7.1 Διαγράμματα 3D περιγραμμάτων	116
5.8 Αποθήκευση και εκτύπωση γραφικών	117
Συμπεράσματα.....	118

• Μέρος Δεύτερο •

Κεφάλαιο 6 - Εισαγωγή στη στατιστική. Περιγραφική στατιστική..... 121

Εισαγωγή.....	121
6.1 Μέτρα κεντρικής τάσης.....	121
6.1.1 Μέση τιμή.....	123
6.1.2 Διάμεση τιμή.....	124
6.2 Μέτρα διασποράς.....	125
6.2.1 Διασπορά (variance).....	125
6.2.2 Τυπική απόκλιση.....	126
6.2.3 Τυπικό σφάλμα μέσης τιμής (standard error of the mean).....	126
6.3 Ιστογράμματα.....	127
6.3.1 Χρήση μεταβλητών ιστογραμμάτων.....	128
6.4 Ποσοστημόρια.....	133
6.5 Θηκογράμματα.....	136
6.5.1 Παραλλαγές θηκογραμμάτων.....	138
6.6 Δειγματοληψία και τυχαία δείγματα.....	140
6.7 Συναρτήσεις προσομοίωσης κατανομών.....	143
6.7.1 Τυχαίοι αριθμοί από ομοιόμορφη κατανομή.....	143
6.7.2 Κανονική κατανομή.....	144
6.7.3 Προσομοίωση διωνυμικών πιθανοτήτων.....	145
6.7.4 Προσομοίωση σπάνιων γεγονότων με την κατανομή Poisson.....	146
Εφαρμογή: Ανάλυση Μαθητικών Επιδόσεων.....	147
Συμπεράσματα.....	161

Κεφάλαιο 7 - Επαγωγή και έλεγχος υποθέσεων 163

Εισαγωγή.....	163
7.1 Έλεγχος κανονικότητας.....	163
7.1.1 Γραφικός έλεγχος κανονικότητας.....	165
7.1.2 Αριθμητικοί έλεγχοι κανονικότητας.....	167
7.2 Σύγκριση μέσων τιμών.....	170
7.2.1 Σε κανονικά κατανεμημένα δείγματα.....	172
7.2.2 Σε μη κανονικά κατανεμημένα δείγματα.....	175
7.3 Σύγκριση λόγων και αναλογιών.....	177

7.3.1 Έλεγχος Fisher για πίνακες σύμπτωσης 2Χ2	178
7.3.2 Έλεγχος για πίνακες σύμπτωσης >2Χ2	179
7.4 Έλεγχος υπερ-εκπροσωπήσεων με την υπεργεωμετρική κατανομή	181
7.5 Στατιστικοί έλεγχοι μέσω μεταθέσεων (permutation tests)	185
Εφαρμογή: Ανάλυση αξίας Ακινήτων στην Πολιτεία της Καλιφόρνια	192
Συμπεράσματα	208

Κεφάλαιο 8 - Ανάλυση διακύμανσης και έλεγχοι Πολλαπλών Υποθέσεων 209

Εισαγωγή	209
8.1 Πολλαπλοί ζευγαρωτοί έλεγχοι (pairwise tests)	209
8.2 Έλεγχος πολλαπλών υποθέσεων	212
8.2.1 Διόρθωση τιμής p-value	214
8.3 Έλεγχος διακύμανσης	215
8.4 Ανάλυση διακύμανσης (ANOVA)	216
8.4.1 Ανάλυση διακύμανσης: στην θεωρία	217
8.4.2 Ανάλυση διακύμανσης: στην πράξη	220
8.5 Πολύ-παραγοντική ANOVA (multi-way ANOVA)	224
8.6 Προϋποθέσεις για την διενέργεια ANOVA	230
8.6.1 Ανεξαρτησία	231
8.6.2 Κανονικότητα	231
8.6.3 Ομοσκεδαστικότητα	232
8.6.4 Ισορροπημένα και μη-ισορροπημένα δείγματα	234
8.7 ANOVA σε μη κανονικά κατανομημένα δείγματα	236
Εφαρμογή: Ανάλυση Επιπέδων Γονιδιακής Έκφρασης	233
Συμπέρασμα	256

Κεφάλαιο 9 - Συσχέτιση και γραμμική παλινδρόμηση 257

Εισαγωγή	257
9.1 Συσχέτιση (correlation)	257
9.1.1 Γραμμική συσχέτιση Pearson (Pearson linear corellation)	258
9.1.2 Συσχέτιση και ελλiptείες τιμές	261
9.1.3 Συσχετίσεις κατάταξης (rank correlations)	262
9.2 Διακύμανση και μερική συσχέτιση (covariance)	265

9.3 Παλινδρόμηση (regression).....	267
9.4 Απλή γραμμική παλινδρόμηση	268
9.5 Παλινδρόμηση με την συνάρτηση $lm()$	269
9.5.1 Απλή παλινδρόμηση.....	269
9.5.2 Πολυωνμική παλινδρόμηση.....	275
9.5.3 Πολλαπλή παλινδρόμηση.....	279
9.6 Στοιχεία διάγνωσης μοντέλων παλινδρόμησης	283
9.6.1 Προσαρμογή.....	283
9.6.2 Κανονικότητα και ακραίες τιμές.....	284
9.6.3 Ομοσκεδαστικότητα και τιμές υψηλής μόχλευσης (high leverage).....	285
9.6.4 Απόσταση Cook και επιδραστικές τιμές.....	287
9.6.5 Συνολική διάγνωση ιδιαίτερων τιμών.....	288
9.7 Σύγκριση μοντέλων και ιεράρχηση παραμέτρων.....	290
Εφαρμογή: Τιμές Lego Sets.	
Πόσο αξίζει στ' αλήθεια το Millenium Falcon;	294
Συμπεράσματα.....	309

• Μέρος Τρίτο •

Κεφάλαιο 10 - Μείωση Διαστασιμότητας.....	313
Εισαγωγή.....	314
10.1 Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis, PCA) ...	314
10.1.1 Σχηματική ερμηνεία της PCA.....	317
10.1.2 Πρακτική εφαρμογή της PCA	317
10.1.3 Γραφική αναπαράσταση PCA.....	319
10.2 Πολυδιάστατη Κλιμάκωση (Multidimensional Scaling, MDS)	322
10.2.1 Πίνακες αποστάσεων	322
10.2.2 Εφαρμογή MDS σε πίνακες αποστάσεων	324
10.3 Στοχαστική Ενσωμάτωση Γειτόνων (t-distributed Stochastic Neighbour Embedding, tSNE).....	327
10.3.1 Αρχή της μεθοδολογίας tSNE	327
10.3.2 Αρχή της μεθοδολογίας tSNE	328
10.4 Διερευνητική Ανάλυση Παραγόντων (Exploratory Factor Analysis, EFA)	332

10.4.1 “Λανθάνουσες” μεταβλητές έναντι γραμμικών/ μη-γραμμικών συνδυασμών	332
10.4.2 Εφαρμογή της EFA	333
10.5 Συμπεράσματα	339
Εφαρμογή: Κυνηγοί Ταλέντων. Πώς να διαλέξουμε τον επόμενο επιθετικό μας;	340
Συμπεράσματα	358

Κεφάλαιο 11 - Ομαδοποίηση 359

Εισαγωγή	359
11.1 Γενικές Αρχές Ομαδοποίησης	360
11.2 Στόχοι της Ομαδοποίησης	361
11.3 Ιεραρχική Ομαδοποίηση (Hierarchical Clustering)	361
11.3.1 Υπολογισμός Αποστάσεων/Ομοιοτήτων	362
11.3.2 Εφαρμογή Ιεραρχικής Ομαδοποίησης	364
11.3.3 Εξαγωγή ομάδων από Ιεραρχικά δέντρα	368
11.3.4 Ιεραρχική Ομαδοποίηση σε Θερμικούς Χάρτες	372
11.4 Ομαδοποίηση Διαμερισμού (Partition Clustering)	377
11.4.1 Υπολογισμός βέλτιστου αριθμού ομάδων	377
11.4.1.1 Συναρτήσεις Αριθμητικών Διανυσμάτων	378
11.4.1.2 Μέτρα συνεκτικότητας ομάδων: Πολλαπλά διαγνωστικά τεστ	379
11.4.2 Ομαδοποίηση κ-μέσων (k-means clustering)	383
11.4.3 Ομαδοποίηση διαμερισμού μεσοειδών (Partitioning Around Medoids, PAM)	389
11.5 Ομαδοποίηση μέσω πυκνότητας	392
11.5.1 Ομαδοποίηση μέσω πυκνότητας με εντοπισμό θορύβου (DBSCAN)	393
11.5.2 Υπολογισμός ακτίνας ϵ και ελάχιστου αριθμού στοιχείων πυρήνα N_{min}	394
11.5.3 Εφαρμογή της μεθόδου DBSCAN	395
11.6 Συμπεράσματα	398
Εφαρμογή: Διάγνωση Καρκίνου του Μαστού. Πότε είμαστε σίγουροι για τα κακά νέα;	398
Συμπεράσματα	418

Κεφάλαιο 12 - Ταξινόμηση

Εισαγωγή.....	419
12.1 Ταξινόμηση, ομαδοποίηση και παλινδρόμηση.....	419
12.2 Εκμάθηση υπό επίβλεψη. Σύνολα εκμάθησης και Σύνολα ελέγχου.....	420
12.3 Μέθοδοι που σχετίζονται με την ομαδοποίηση. Ταξινόμηση k πλησιέστερων γειτόνων (k Nearest Neighbours, kNN)	421
12.3.1 Προετοιμασία Δεδομένων.....	422
12.3.2 Εφαρμογή kNN.....	423
12.3.3 Επιλογή παραμέτρων για kNN	424
12.4 Δυαδική Ταξινόμηση με Λογιστική Παλινδρόμηση (Logistic Regression).....	427
12.4.1 Εκμάθηση της λογιστικής παλινδρόμησης.....	428
12.4.2 Εφαρμογή λογιστικής παλινδρόμησης. Η συνάρτηση <i>predict()</i>	430
12.4.3 Αξιολόγηση λογιστικής παλινδρόμησης.....	432
12.5 Γενικευμένα Γραμμικά μοντέλα για μη-κανονικές κατανομές (Generalized Linear Models, GLM).....	436
12.6 Δέντρα αποφάσεων (Decision Trees).....	440
12.6.1 Θεωρητική βάση δέντρων αποφάσεων	440
12.6.2 Εφαρμογή και αξιολόγηση δέντρων αποφάσεων.....	444
12.7 Τυχαία Δάση (Random Forests).....	450
12.7.1 Τυχαία Δάση. Εφαρμογή	450
12.7.2 Ιεράρχηση επεξηγηματικών μεταβλητών στα Τυχαία Δάση	452
12.8 Μηχανές Διανυσματικής Στήριξης (Support Vector Machines, SVM).....	454
12.8.1 Βασικές αρχές των SVM	454
12.8.2 Εφαρμογή των SVM.....	455
12.8.3 Ρύθμιση παραμέτρων SVM (SVM tuning).....	456
12.9 Συμπεράσματα.....	458
Εφαρμογή: Ταξινόμηση πελατών. Σε ποιους θα πρέπει να κάνουμε έκπτωση;.....	459
Συμπεράσματα.....	479