

# Contents

<b>1</b>	<b>The Art of Analyzing Data</b>	<b>3</b>
1.1	What this book is about . . . . .	5
1.1.1	Useful data characterizations . . . . .	6
1.1.2	Ohm's law . . . . .	8
1.2	How much can we learn from data? . . . . .	11
1.2.1	Can one hear the shape of a drum? . . . . .	11
1.2.2	The role of assumptions . . . . .	12
1.3	Numerical mathematics versus data analysis . . . . .	13
1.3.1	Numbers, arithmetic, and roundoff errors . . . . .	14
1.3.2	Computing mathematical functions . . . . .	16
1.3.3	Data analysis and uncertainty . . . . .	17
1.4	Dealing with uncertain data . . . . .	17
1.4.1	Additive uncertainty models . . . . .	18
1.4.2	The minimum uncertainty model . . . . .	19
1.4.3	The random variable model . . . . .	20
1.4.4	Other uncertainty models . . . . .	22
1.5	What is a good data model? . . . . .	24
1.5.1	Empirical versus fundamental models . . . . .	24
1.5.2	The principle of zig-zag-and-swirl . . . . .	26
1.5.3	Ockham's razor and overfitting . . . . .	27
1.5.4	Wei's elephant and Einstein's advice . . . . .	31
1.6	Exploratory versus confirmatory analysis . . . . .	32
1.6.1	Confirmatory data analysis . . . . .	32
1.6.2	Exploratory data analysis . . . . .	34
1.6.3	A cautionary example: the killer potato . . . . .	34
1.7	The four R's of exploratory data analysis . . . . .	37
1.7.1	The first R: revelation . . . . .	38
1.7.2	The second R: residuals . . . . .	40
1.7.3	The third R: reexpression . . . . .	42
1.7.4	The fourth R: resistance . . . . .	45
1.8	Working with real datasets . . . . .	48
1.8.1	A missing data example: the asteroid belt . . . . .	48
1.8.2	A large dataset: chronic fatigue syndrome . . . . .	50
1.9	Software considerations . . . . .	52

1.10	Organization of the rest of this book . . . . .	53
<b>2</b>	<b>Data: Types, Uncertainty and Quality</b>	<b>56</b>
2.1	The structure of datasets . . . . .	58
2.2	Data types . . . . .	59
2.3	Metadata . . . . .	62
2.4	What can we measure and how well? . . . . .	65
2.4.1	What can we measure? . . . . .	66
2.4.2	Accuracy and precision . . . . .	72
2.4.3	The limits of measurement quality . . . . .	73
2.4.4	More typical measurements . . . . .	75
2.5	Variations: normal and anomalous . . . . .	80
2.5.1	Normal variations and “noise” . . . . .	80
2.5.2	Outliers: gross errors and legitimate surprises . . . . .	82
2.5.3	Inliers: a subtle data anomaly . . . . .	86
2.6	Missing data . . . . .	87
2.6.1	The problem of coding missing data . . . . .	88
2.6.2	Disguised missing data . . . . .	90
2.6.3	Causes of missing data . . . . .	93
2.6.4	Ignorable versus nonignorable missing data . . . . .	95
2.7	Other data anomalies . . . . .	96
2.7.1	Coarse quantization . . . . .	96
2.7.2	Noninformative variables . . . . .	97
2.7.3	File merge and manipulation errors . . . . .	99
2.7.4	Duplicate records . . . . .	101
2.7.5	Categorical data errors . . . . .	103
2.8	A few concluding observations . . . . .	103
<b>3</b>	<b>Characterizing Categorical Variables</b>	<b>106</b>
3.1	Three categorical data examples . . . . .	108
3.1.1	The UCI mushroom dataset . . . . .	108
3.1.2	Who wrote the <i>Federalist Papers</i> ? . . . . .	109
3.1.3	Horse-kick deaths in the Prussian army . . . . .	110
3.2	Discrete random variables . . . . .	113
3.2.1	The discrete random variable model . . . . .	113
3.2.2	Events and probabilities . . . . .	115
3.3	Three important distributions . . . . .	117
3.3.1	Urn models and the binomial distribution . . . . .	117
3.3.2	The hypergeometric distribution . . . . .	120
3.3.3	The discrete uniform distribution . . . . .	123
3.4	Entropy . . . . .	123
3.5	Interestingness and heterogeneity . . . . .	125
3.5.1	Four heterogeneity measures . . . . .	126
3.5.2	Application to the UCI mushroom dataset . . . . .	128
3.6	Count distributions . . . . .	132
3.6.1	The Poisson distribution . . . . .	132

3.6.2	The negative binomial distribution . . . . .	133
3.6.3	Zero-inflated count models . . . . .	134
3.7	The Zipf distribution . . . . .	135
3.7.1	Definition and properties . . . . .	135
3.7.2	Examples and consequences . . . . .	136
3.8	Exercises . . . . .	138
<b>4</b>	<b>Uncertainty in Real Variables</b>	<b>139</b>
4.1	Continuous random variables . . . . .	141
4.1.1	Distributions and densities . . . . .	141
4.1.2	Location parameters: mean, median, and mode . . . . .	144
4.1.3	Expected values and moments . . . . .	146
4.2	How are data values distributed? . . . . .	148
4.2.1	The normal (Gaussian) distribution . . . . .	148
4.2.2	Clancey’s survey of data distributions . . . . .	150
4.3	Moment characterizations . . . . .	152
4.3.1	The Markov and Chebyshev inequalities . . . . .	152
4.3.2	Skewness and kurtosis . . . . .	154
4.3.3	The method of moments . . . . .	157
4.3.4	Karl Pearson’s 1895 system of distributions . . . . .	158
4.3.5	Johnson’s system of distributions . . . . .	159
4.4	Limitations of moment characterizations . . . . .	161
4.4.1	Exact characterizations . . . . .	161
4.4.2	Approximate characterizations . . . . .	162
4.5	Some important distributions . . . . .	166
4.5.1	The beta distribution . . . . .	166
4.5.2	The Cauchy distribution . . . . .	169
4.5.3	The exponential distribution . . . . .	171
4.5.4	The gamma distribution . . . . .	172
4.5.5	The Laplace distribution . . . . .	173
4.5.6	The logistic distribution . . . . .	175
4.5.7	The lognormal distribution . . . . .	178
4.5.8	The Pareto distribution . . . . .	180
4.5.9	The Rayleigh distribution . . . . .	181
4.5.10	The Weibull distribution . . . . .	182
4.6	Exercises . . . . .	184
<b>5</b>	<b>Fitting Straight Lines</b>	<b>186</b>
5.1	Why do we fit straight lines? . . . . .	188
5.1.1	Linear constitutive relations . . . . .	188
5.1.2	Taylor series expansions . . . . .	190
5.1.3	Allometry . . . . .	192
5.1.4	Behavior and functional equations . . . . .	193
5.2	Do we fit $\mathbf{y}$ on $\mathbf{x}$ or $\mathbf{x}$ on $\mathbf{y}$ ? . . . . .	195
5.3	Three approaches to fitting lines . . . . .	197
5.3.1	Optimization-based problem formulations . . . . .	198

5.3.2	The ordinary least squares (OLS) fit . . . . .	200
5.3.3	The least absolute deviations (LAD) fit . . . . .	201
5.3.4	The total least squares (TLS) fit . . . . .	202
5.4	The method of maximum likelihood . . . . .	205
5.4.1	The basic concept . . . . .	206
5.4.2	Three specific maximum likelihood solutions . . . . .	207
5.5	Two brief case studies . . . . .	210
5.5.1	Case study 1: $\mathbf{L}_1$ vs. $\mathbf{L}_2$ vs. $\mathbf{L}_\infty$ . . . . .	210
5.5.2	Case study 2: OLS vs. TLS . . . . .	214
5.6	The unknown-but-bounded formulation . . . . .	217
5.7	Which method do we use? . . . . .	221
5.8	Exercises . . . . .	223
<b>6</b>	<b>A Brief Introduction to Estimation Theory</b>	<b>225</b>
6.1	Characterizing estimators . . . . .	227
6.1.1	Location estimators . . . . .	227
6.1.2	Estimator bias . . . . .	234
6.1.3	Variance and consistency . . . . .	235
6.1.4	Other characterizations . . . . .	236
6.2	An example: variance estimation . . . . .	237
6.2.1	The standard estimators $\bar{x}_N$ and $\hat{\sigma}^2$ . . . . .	237
6.2.2	Exact distribution for the Gaussian case . . . . .	238
6.2.3	What about non-Gaussian cases? . . . . .	239
6.3	The CLT and asymptotic normality . . . . .	241
6.3.1	Distributions of sums and averages . . . . .	242
6.3.2	The Central Limit Theorem . . . . .	246
6.3.3	Asymptotic normality and relative efficiency . . . . .	249
6.4	Cases where the CLT does not apply . . . . .	251
6.4.1	Stable random variables . . . . .	251
6.4.2	Weighted averages . . . . .	253
6.4.3	Webster's ambient noise statistics . . . . .	255
6.5	The information inequality . . . . .	256
6.6	Order statistics and L-estimators . . . . .	259
6.6.1	Characterizing the Cauchy distribution . . . . .	261
6.6.2	Distributions of order statistics . . . . .	261
6.6.3	Uniform order statistics . . . . .	263
6.6.4	A maximum likelihood estimation problem . . . . .	265
6.6.5	L-estimators and their properties . . . . .	267
6.6.6	L-estimators for Cauchy parameters . . . . .	268
6.6.7	Gastwirth's location estimator . . . . .	269
6.6.8	Asymptotic normality of L-estimators . . . . .	269
6.6.9	Gini's mean difference . . . . .	271
6.6.10	Uniform maximum likelihood estimators . . . . .	271
6.7	Exercises . . . . .	273

<b>7</b>	<b>Outliers: Distributional Monsters (?) That Lurk in Data</b>	<b>275</b>
7.1	Outliers and their consequences . . . . .	277
7.1.1	The outlier sensitivity of moments . . . . .	277
7.1.2	Failure of the $3\sigma$ -edit rule . . . . .	280
7.1.3	The contaminated normal outlier model . . . . .	286
7.2	Four ways of dealing with outliers . . . . .	287
7.2.1	Detect and omit . . . . .	288
7.2.2	Detect and replace . . . . .	290
7.2.3	Detect and scrutinize . . . . .	291
7.2.4	Use outlier-resistant analytical procedures . . . . .	294
7.3	Robust estimators . . . . .	295
7.3.1	The breakdown point: a measure of resistance . . . . .	296
7.3.2	The influence function: a measure of smoothness . . . . .	297
7.3.3	Efficiency robustness: a measure of breadth . . . . .	298
7.3.4	A comparison of the mean and the median . . . . .	299
7.4	Robust alternatives to $\bar{x}_N$ and $\hat{\sigma}$ . . . . .	300
7.4.1	The Princeton robustness study . . . . .	300
7.4.2	The MADM scale estimate . . . . .	302
7.4.3	Robustness of the MADM scale estimate . . . . .	304
7.5	Outlier detection . . . . .	305
7.5.1	The Hampel identifier . . . . .	305
7.5.2	Masking and swamping breakdown points . . . . .	308
7.5.3	Practical details in outlier detection . . . . .	310
7.6	The problem of asymmetry . . . . .	312
7.6.1	Robust asymmetry measures . . . . .	312
7.6.2	Location-free scale estimates . . . . .	316
7.7	Other practical considerations . . . . .	318
7.7.1	Light-tailed and bimodal distributions . . . . .	318
7.7.2	Discrete distributions (quantization) . . . . .	319
7.7.3	Discontinuity—a cautionary tale . . . . .	322
7.8	General recommendations . . . . .	324
7.8.1	How robust is enough? . . . . .	324
7.8.2	Overall recommendations . . . . .	325
7.9	Exercises . . . . .	327
<b>8</b>	<b>Characterizing a Dataset</b>	<b>329</b>
8.1	Surveying and appreciating a dataset . . . . .	331
8.2	Three useful visualization tools . . . . .	332
8.2.1	The normal Q-Q plot . . . . .	332
8.2.2	The Poissonness plot . . . . .	334
8.2.3	Nonparametric density estimators . . . . .	335
8.3	Quantile-quantile plots . . . . .	337
8.3.1	The basic idea . . . . .	337
8.3.2	The general construction . . . . .	339
8.3.3	Normal Q-Q plots . . . . .	341
8.3.4	Data comparison plots . . . . .	346

8.4	Plots for discrete distributions . . . . .	348
8.4.1	Poissonness plots . . . . .	348
8.4.2	Negative binomialness plots . . . . .	350
8.5	Histograms: crude density estimates . . . . .	355
8.5.1	The basic histogram . . . . .	355
8.5.2	Histogram bias . . . . .	357
8.5.3	Histogram variance . . . . .	359
8.6	Kernel density estimators . . . . .	360
8.6.1	The basic kernel estimator . . . . .	361
8.6.2	Bias in kernel estimates . . . . .	363
8.6.3	Variance of kernel estimates . . . . .	364
8.6.4	A comparison of four examples . . . . .	365
8.7	Scatterplot smoothers . . . . .	370
8.7.1	<b>Supsmu</b> : an adaptive smoother . . . . .	373
8.7.2	The lowess smoother . . . . .	375
8.8	The preliminary data survey . . . . .	378
8.9	Exercises . . . . .	383
<b>9</b>	<b>Confidence Intervals and Hypothesis Testing</b>	<b>386</b>
9.1	Confidence intervals . . . . .	388
9.1.1	Application: systematic errors . . . . .	390
9.1.2	The case of unknown variance . . . . .	392
9.2	Extensions of the Poissonness plot . . . . .	393
9.3	Formal hypothesis tests . . . . .	394
9.4	Comparing means . . . . .	396
9.4.1	The classical <b>t</b> -test . . . . .	397
9.4.2	Limitations of the <b>t</b> -test . . . . .	398
9.4.3	The Wilcoxon rank-sum test . . . . .	400
9.4.4	The Yuen-Welch and Welch rank-based tests . . . . .	403
9.5	The $\chi^2$ distribution and $\chi^2$ tests . . . . .	406
9.5.1	The $\chi^2$ distribution . . . . .	406
9.5.2	The $\chi^2$ test . . . . .	407
9.5.3	An application: uniformity testing . . . . .	408
9.6	The <b>F</b> -test . . . . .	410
9.7	Binomial random variables . . . . .	412
9.8	Testing multiple hypotheses . . . . .	415
9.8.1	The multiple comparison problem . . . . .	415
9.8.2	The Bonferroni correction . . . . .	416
9.8.3	The Holm stepdown procedure . . . . .	417
9.8.4	The Benjamani-Hochberg procedure . . . . .	418
9.9	Exercises . . . . .	419

<b>10 Relations among Variables</b>	<b>421</b>
10.1 What is the relationship between popular and electoral votes? . . .	423
10.1.1 Analysis of the <i>World Almanac</i> data . . . . .	423
10.1.2 Electoral versus popular vote margins . . . . .	426
10.1.3 Association measures . . . . .	428
10.1.4 Data limitations and anomalies . . . . .	429
10.2 Joint and conditional distributions . . . . .	430
10.2.1 Discrete events: the multinomial distribution . . . . .	431
10.2.2 Multivariate distributions and densities . . . . .	433
10.2.3 Statistical independence . . . . .	435
10.2.4 Conditional probabilities . . . . .	436
10.2.5 Conditional distributions and expectations . . . . .	437
10.3 The multivariate Gaussian distribution . . . . .	440
10.3.1 Vector formulation . . . . .	441
10.3.2 Mahalanobis distances . . . . .	443
10.3.3 The bivariate case . . . . .	444
10.3.4 Quadrant probabilities . . . . .	449
10.4 The product-moment correlation coefficient . . . . .	450
10.4.1 Definition and estimation . . . . .	451
10.4.2 Exact distribution for the Gaussian case . . . . .	452
10.4.3 Fisher’s transformation to normality . . . . .	454
10.4.4 Testing for independence . . . . .	455
10.4.5 The influence of outliers . . . . .	456
10.4.6 The influence of transformations . . . . .	459
10.5 The Spearman rank correlation coefficient . . . . .	462
10.5.1 Definition, estimation, and properties . . . . .	462
10.5.2 The influence of outliers . . . . .	464
10.5.3 An application of rank correlations . . . . .	465
10.6 Mixture distributions . . . . .	469
10.6.1 Discrete mixtures . . . . .	469
10.6.2 Example: Gaussian mixtures . . . . .	472
10.6.3 Continuous mixtures . . . . .	476
10.6.4 Example 1: overdispersion . . . . .	477
10.6.5 Example 2: ratios of random variables . . . . .	479
10.6.6 Example 3: heavy-tailed distributions . . . . .	480
10.7 Non-Gaussian multivariate distributions . . . . .	482
10.7.1 Bivariate exponential distributions . . . . .	483
10.7.2 Two surprising “near-Gaussian” examples . . . . .	484
10.7.3 Elliptically distributed random variables . . . . .	486
10.7.4 Copulas: building from marginals . . . . .	487
10.7.5 Kendall’s $\tau$ . . . . .	490
10.8 Relations among other variable types . . . . .	491
10.8.1 Discrete, ordinal, and nominal variables . . . . .	491
10.8.2 Mixed data types . . . . .	492
10.8.3 The special case of binary variables . . . . .	493
10.9 Exercises . . . . .	494

<b>11 Regression Models I: Real Data</b>	<b>498</b>
11.1 Building regression models . . . . .	500
11.1.1 Linear versus nonlinear regression . . . . .	500
11.1.2 An example: the Riedel equation . . . . .	502
11.1.3 A second example: dimensional analysis . . . . .	503
11.2 Ordinary least squares (OLS) . . . . .	504
11.3 Two simple OLS extensions . . . . .	508
11.3.1 Weighted least squares . . . . .	509
11.3.2 Restricted least squares . . . . .	510
11.4 M-estimators and robust regression . . . . .	511
11.4.1 Basic notions of M-estimators . . . . .	511
11.4.2 Mechanics of M-estimators . . . . .	513
11.4.3 Two illustrative examples . . . . .	514
11.5 Other robust alternatives to OLS . . . . .	516
11.6 Exercises . . . . .	519
<b>12 Reexpression: Data Transformations</b>	<b>521</b>
12.1 Three uses for transformations . . . . .	523
12.1.1 Changing visual emphasis . . . . .	523
12.1.2 Linearizing nonlinear models . . . . .	525
12.1.3 Changing data distributions . . . . .	527
12.2 Four transformation horror stories . . . . .	529
12.2.1 Making unrelated variables appear related . . . . .	530
12.2.2 Transformations need not preserve curvature . . . . .	534
12.2.3 Transformations need not preserve modality . . . . .	536
12.2.4 “Everything looks linear on a log-log plot” . . . . .	537
12.3 Three popular transformations . . . . .	541
12.3.1 Box-Cox transformations . . . . .	541
12.3.2 Aranda-Ordaz transformations . . . . .	542
12.3.3 The angular transformation . . . . .	544
12.4 Characteristics of good transformations . . . . .	545
12.5 Generating nonuniform random numbers . . . . .	548
12.5.1 Exponentially distributed random samples . . . . .	548
12.5.2 Cauchy distributed random samples . . . . .	548
12.5.3 Logistic distributed random samples . . . . .	549
12.5.4 Pareto distributed random samples . . . . .	549
12.5.5 Weibull distributed random samples . . . . .	549
12.6 More general transformations . . . . .	550
12.6.1 Transforming densities . . . . .	550
12.6.2 Transformed exponential random variables . . . . .	551
12.6.3 The $\chi_1^2$ density . . . . .	553
12.7 Reciprocal transformations . . . . .	554
12.7.1 The Gaussian distribution . . . . .	555
12.7.2 The Laplace distribution . . . . .	557
12.7.3 The Cauchy distribution . . . . .	558
12.7.4 The beta and Pareto distributions . . . . .	558



12.7.5	The lognormal distribution . . . . .	559
12.8	Exercises . . . . .	559
<b>13</b>	<b>Regression Models II: Mixed Data Types</b>	<b>561</b>
13.1	Models with mixed data types . . . . .	563
13.2	The influences of data type . . . . .	564
13.2.1	Binary association: the odds ratio . . . . .	564
13.2.2	Do big animals have big brains? . . . . .	565
13.3	ANOVA models . . . . .	567
13.3.1	Analysis of variance (ANOVA) . . . . .	567
13.3.2	Extensions and practical issues . . . . .	571
13.3.3	Application: bitter pit in apples . . . . .	577
13.4	Generalized linear models . . . . .	581
13.5	Logistic regression models . . . . .	585
13.5.1	Logistic regression . . . . .	585
13.5.2	Ungrouped data . . . . .	587
13.5.3	Application 1: bitter pit revisited . . . . .	588
13.5.4	Application 2: missing data . . . . .	592
13.5.5	Practical issues: EPV and separation . . . . .	595
13.6	Poisson regression models . . . . .	597
13.6.1	Over- and underdispersion . . . . .	597
13.6.2	Poisson regression . . . . .	599
13.6.3	Application: NPG in the Pima Indians dataset . . . . .	600
13.7	Exercises . . . . .	603
<b>14</b>	<b>Characterizing Analysis Results</b>	<b>605</b>
14.1	Analyzing modified datasets . . . . .	607
14.1.1	Variation-based analysis procedures . . . . .	607
14.1.2	The notion of exchangeability . . . . .	608
14.2	Computational negative controls . . . . .	610
14.2.1	Empirical probabilities and <b>z</b> -scores . . . . .	612
14.2.2	An application: assessing correlations . . . . .	613
14.3	Deletion diagnostics . . . . .	616
14.3.1	The deletion diagnostic framework . . . . .	616
14.3.2	Simulation example: correlation analysis . . . . .	617
14.3.3	Application to the brain/body dataset . . . . .	620
14.4	Bootstrap resampling methods . . . . .	622
14.4.1	The basic bootstrap formulation . . . . .	623
14.4.2	Application to the brain/body dataset . . . . .	623
14.5	Subsampling methods . . . . .	625
14.5.1	Subsampling versus the bootstrap . . . . .	625
14.5.2	Application 1: the simulation dataset . . . . .	626
14.5.3	Application 2: the brain/body dataset . . . . .	628
14.6	Applicability of these methods . . . . .	630
14.7	Exercises . . . . .	635

<b>15 Regression Models III: Diagnostics and Refinements</b>	<b>636</b>
15.1 The model-building process . . . . .	638
15.1.1 What is a good data model? . . . . .	638
15.1.2 The model development cycle . . . . .	640
15.2 Three modeling examples . . . . .	641
15.2.1 The brain/body dataset . . . . .	641
15.2.2 Predicting triceps skinfold thickness . . . . .	642
15.2.3 Bitter pit and mineral content . . . . .	643
15.3 Assessing goodness-of-fit . . . . .	643
15.3.1 The classical $\mathbf{R}^2$ measure and $\mathbf{F}$ -statistics . . . . .	644
15.3.2 Robust goodness-of-fit measures . . . . .	646
15.4 Initial variable selection . . . . .	648
15.4.1 Statistical significance versus purposeful selection . . . . .	648
15.4.2 Considering variable transformations . . . . .	649
15.5 Deciding which variables to keep . . . . .	654
15.6 The problem of collinearity . . . . .	657
15.6.1 Collinearity and OLS parameter estimates . . . . .	658
15.6.2 Dealing with collinearity . . . . .	660
15.7 Finding influential data observations . . . . .	662
15.7.1 Examining model residuals . . . . .	662
15.7.2 Leverage and the hat matrix . . . . .	664
15.7.3 OLS regression diagnostics . . . . .	666
15.8 Cross-validation . . . . .	669
15.8.1 Leave-one-out cross-validation . . . . .	671
15.8.2 $\mathbf{K}$ -fold cross-validation . . . . .	672
15.8.3 Cross-validation for variable selection . . . . .	673
15.9 Iterative refinement strategies . . . . .	675
15.9.1 All subsets regression . . . . .	676
15.9.2 Stepwise regression . . . . .	677
15.9.3 Forward selection for the TSF model . . . . .	678
15.9.4 An alternative TSF model . . . . .	678
15.9.5 Forward selection for the bitter pit model . . . . .	680
15.10 Exercises . . . . .	681
<b>16 Dealing with Missing Data</b>	<b>684</b>
16.1 The missing data problem . . . . .	686
16.1.1 General missing data strategies . . . . .	686
16.1.2 Missingness: MCAR, MAR, and MNAR . . . . .	688
16.2 The univariate case . . . . .	689
16.2.1 Univariate issues and strategies . . . . .	690
16.2.2 Four simulation examples . . . . .	691
16.2.3 Location estimates . . . . .	691
16.2.4 Scale estimates . . . . .	696
16.3 Four multivariate examples . . . . .	700
16.4 Case deletion strategies . . . . .	705
16.4.1 Complete case versus available case analysis . . . . .	705

16.4.2	Omitting variables and related ideas . . . . .	706
16.4.3	Results for the simulation examples . . . . .	707
16.5	Simple imputation strategies . . . . .	710
16.5.1	Mean imputation . . . . .	710
16.5.2	Hot-deck imputation . . . . .	712
16.5.3	Regression-based imputation . . . . .	715
16.5.4	Hot-deck/regression composite method . . . . .	718
16.6	Multiple imputation . . . . .	720
16.7	The EM algorithm . . . . .	723
16.7.1	General description . . . . .	723
16.7.2	A simple univariate example . . . . .	724
16.7.3	A nonignorable univariate example . . . . .	726
16.7.4	Specialization to bivariate Gaussian data . . . . .	728
16.7.5	Application to the simulation examples . . . . .	731
16.7.6	The Healy-Westmacott regression procedure . . . . .	733
16.8	Results for the Pima Indians dataset . . . . .	735
16.9	General conclusions . . . . .	738
16.10	Exercises . . . . .	740
<b>Bibliography</b>		<b>745</b>
<b>Index</b>		<b>765</b>