

INTERNATIONAL GEOPHYSICS SERIES

STATISTICAL METHODS IN THE ATMOSPHERIC SCIENCES

Second Edition



DANIEL S. WILKS

Contents

<i>Preface to the First Edition</i>	xv
<i>Preface to the Second Edition</i>	xvii
PART I Preliminaries	1
CHAPTER 1 Introduction	3
1.1 What Is Statistics?	3
1.2 Descriptive and Inferential Statistics	3
1.3 Uncertainty about the Atmosphere	4
CHAPTER 2 Review of Probability	7
2.1 Background	7
2.2 The Elements of Probability	7
2.2.1 Events	7
2.2.2 The Sample Space	8
2.2.3 The Axioms of Probability	9
2.3 The Meaning of Probability	9
2.3.1 Frequency Interpretation	10
2.3.2 Bayesian (Subjective) Interpretation	10
2.4 Some Properties of Probability	11
2.4.1 Domain, Subsets, Complements, and Unions	11
2.4.2 DeMorgan's Laws	13
2.4.3 Conditional Probability	13
2.4.4 Independence	14
2.4.5 Law of Total Probability	16
2.4.6 Bayes' Theorem	17
2.5 Exercises	18

PART II Univariate Statistics 21

CHAPTER 3 Empirical Distributions and Exploratory Data Analysis 23

- 3.1 Background 23
 - 3.1.1 Robustness and Resistance 23
 - 3.1.2 Quantiles 24
- 3.2 Numerical Summary Measures 25
 - 3.2.1 Location 26
 - 3.2.2 Spread 26
 - 3.2.3 Symmetry 28
- 3.3 Graphical Summary Techniques 28
 - 3.3.1 Stem-and-Leaf Display 29
 - 3.3.2 Boxplots 30
 - 3.3.3 Schematic Plots 31
 - 3.3.4 Other Boxplot Variants 33
 - 3.3.5 Histograms 33
 - 3.3.6 Kernel Density Smoothing 35
 - 3.3.7 Cumulative Frequency Distributions 39
- 3.4 Reexpression 42
 - 3.4.1 Power Transformations 42
 - 3.4.2 Standardized Anomalies 47
- 3.5 Exploratory Techniques for Paired Data 49
 - 3.5.1 Scatterplots 49
 - 3.5.2 Pearson (Ordinary) Correlation 50
 - 3.5.3 Spearman Rank Correlation and Kendall's τ 55
 - 3.5.4 Serial Correlation 57
 - 3.5.5 Autocorrelation Function 58
- 3.6 Exploratory Techniques for Higher-Dimensional Data 59
 - 3.6.1 The Star Plot 59
 - 3.6.2 The Glyph Scatterplot 60
 - 3.6.3 The Rotating Scatterplot 62
 - 3.6.4 The Correlation Matrix 63
 - 3.6.5 The Scatterplot Matrix 65
 - 3.6.6 Correlation Maps 67
- 3.7 Exercises 69

CHAPTER 4 Parametric Probability Distributions 71

- 4.1 Background 71
 - 4.1.1 Parametric vs. Empirical Distributions 71
 - 4.1.2 What Is a Parametric Distribution? 72
 - 4.1.3 Parameters vs. Statistics 72
 - 4.1.4 Discrete vs. Continuous Distributions 73
- 4.2 Discrete Distributions 73
 - 4.2.1 Binomial Distribution 73
 - 4.2.2 Geometric Distribution 76
 - 4.2.3 Negative Binomial Distribution 77
 - 4.2.4 Poisson Distribution 80

- 4.3 Statistical Expectations 82
 - 4.3.1 Expected Value of a Random Variable 82
 - 4.3.2 Expected Value of a Function of a Random Variable 83
- 4.4 Continuous Distributions 85
 - 4.4.1 Distribution Functions and Expected Values 85
 - 4.4.2 Gaussian Distributions 88
 - 4.4.3 Gamma Distributions 95
 - 4.4.4 Beta Distributions 102
 - 4.4.5 Extreme-Value Distributions 104
 - 4.4.6 Mixture Distributions 109
- 4.5 Qualitative Assessments of the Goodness of Fit 111
 - 4.5.1 Superposition of a Fitted Parametric Distribution and Data Histogram 111
 - 4.5.2 Quantile-Quantile (Q-Q) Plots 113
- 4.6 Parameter Fitting Using Maximum Likelihood 114
 - 4.6.1 The Likelihood Function 114
 - 4.6.2 The Newton-Raphson Method 116
 - 4.6.3 The EM Algorithm 117
 - 4.6.4 Sampling Distribution of Maximum-Likelihood Estimates 120
- 4.7 Statistical Simulation 120
 - 4.7.1 Uniform Random Number Generators 121
 - 4.7.2 Nonuniform Random Number Generation by Inversion 123
 - 4.7.3 Nonuniform Random Number Generation by Rejection 124
 - 4.7.4 Box-Muller Method for Gaussian Random Number Generation 126
 - 4.7.5 Simulating from Mixture Distributions and Kernel Density Estimates 127
- 4.8 Exercises 128

CHAPTER 5 Hypothesis Testing 131

- 5.1 Background 131
 - 5.1.1 Parametric vs. Nonparametric Tests 131
 - 5.1.2 The Sampling Distribution 132
 - 5.1.3 The Elements of Any Hypothesis Test 132
 - 5.1.4 Test Levels and p Values 133
 - 5.1.5 Error Types and the Power of a Test 133
 - 5.1.6 One-Sided vs. Two-Sided Tests 134
 - 5.1.7 Confidence Intervals: Inverting Hypothesis Tests 135
- 5.2 Some Parametric Tests 138
 - 5.2.1 One-Sample t Test 138
 - 5.2.2 Tests for Differences of Mean under Independence 140
 - 5.2.3 Tests for Differences of Mean for Paired Samples 141
 - 5.2.4 Test for Differences of Mean under Serial Dependence 143
 - 5.2.5 Goodness-of-Fit Tests 146
 - 5.2.6 Likelihood Ratio Test 154
- 5.3 Nonparametric Tests 156
 - 5.3.1 Classical Nonparametric Tests for Location 156
 - 5.3.2 Introduction to Resampling Tests 162

- 5.3.3 Permutation Tests 164
- 5.3.4 The Bootstrap 166
- 5.4 Field Significance and Multiplicity 170
 - 5.4.1 The Multiplicity Problem for Independent Tests 171
 - 5.4.2 Field Significance Given Spatial Correlation 172
- 5.5 Exercises 176

CHAPTER 6 Statistical Forecasting 179

- 6.1 Background 179
- 6.2 Linear Regression 180
 - 6.2.1 Simple Linear Regression 180
 - 6.2.2 Distribution of the Residuals 182
 - 6.2.3 The Analysis of Variance Table 184
 - 6.2.4 Goodness-of-Fit Measures 185
 - 6.2.5 Sampling Distributions of the Regression Coefficients 187
 - 6.2.6 Examining Residuals 189
 - 6.2.7 Prediction Intervals 194
 - 6.2.8 Multiple Linear Regression 197
 - 6.2.9 Derived Predictor Variables in Multiple Regression 198
- 6.3 Nonlinear Regression 201
 - 6.3.1 Logistic Regression 201
 - 6.3.2 Poisson Regression 205
- 6.4 Predictor Selection 207
 - 6.4.1 Why Is Careful Predictor Selection Important? 207
 - 6.4.2 Screening Predictors 209
 - 6.4.3 Stopping Rules 212
 - 6.4.4 Cross Validation 215
- 6.5 Objective Forecasts Using Traditional Statistical Methods 217
 - 6.5.1 Classical Statistical Forecasting 217
 - 6.5.2 Perfect Prog and MOS 220
 - 6.5.3 Operational MOS Forecasts 226
- 6.6 Ensemble Forecasting 229
 - 6.6.1 Probabilistic Field Forecasts 229
 - 6.6.2 Stochastic Dynamical Systems in Phase Space 229
 - 6.6.3 Ensemble Forecasts 232
 - 6.6.4 Choosing Initial Ensemble Members 233
 - 6.6.5 Ensemble Average and Ensemble Dispersion 234
 - 6.6.6 Graphical Display of Ensemble Forecast Information 236
 - 6.6.7 Effects of Model Errors 242
 - 6.6.8 Statistical Postprocessing: Ensemble MOS 243
- 6.7 Subjective Probability Forecasts 245
 - 6.7.1 The Nature of Subjective Forecasts 245
 - 6.7.2 The Subjective Distribution 246
 - 6.7.3 Central Credible Interval Forecasts 248
 - 6.7.4 Assessing Discrete Probabilities 250
 - 6.7.5 Assessing Continuous Distributions 251
- 6.8 Exercises 252

CHAPTER 7 Forecast Verification 255

- 7.1 Background 255
 - 7.1.1 Purposes of Forecast Verification 255
 - 7.1.2 The Joint Distribution of Forecasts and Observations 256
 - 7.1.3 Scalar Attributes of Forecast Performance 258
 - 7.1.4 Forecast Skill 259
- 7.2 Nonprobabilistic Forecasts of Discrete Predictands 260
 - 7.2.1 The 2×2 Contingency Table 260
 - 7.2.2 Scalar Attributes Characterizing 2×2 Contingency Tables 262
 - 7.2.3 Skill Scores for 2×2 Contingency Tables 265
 - 7.2.4 Which Score? 268
 - 7.2.5 Conversion of Probabilistic to Nonprobabilistic Forecasts 269
 - 7.2.6 Extensions for Multicategory Discrete Predictands 271
- 7.3 Nonprobabilistic Forecasts of Continuous Predictands 276
 - 7.3.1 Conditional Quantile Plots 277
 - 7.3.2 Scalar Accuracy Measures 278
 - 7.3.3 Skill Scores 280
- 7.4 Probability Forecasts of Discrete Predictands 282
 - 7.4.1 The Joint Distribution for Dichotomous Events 282
 - 7.4.2 The Brier Score 284
 - 7.4.3 Algebraic Decomposition of the Brier Score 285
 - 7.4.4 The Reliability Diagram 287
 - 7.4.5 The Discrimination Diagram 293
 - 7.4.6 The ROC Diagram 294
 - 7.4.7 Hedging, and Strictly Proper Scoring Rules 298
 - 7.4.8 Probability Forecasts for Multiple-Category Events 299
- 7.5 Probability Forecasts for Continuous Predictands 302
 - 7.5.1 Full Continuous Forecast Probability Distributions 302
 - 7.5.2 Central Credible Interval Forecasts 303
- 7.6 Nonprobabilistic Forecasts of Fields 304
 - 7.6.1 General Considerations for Field Forecasts 304
 - 7.6.2 The S1 Score 306
 - 7.6.3 Mean Squared Error 307
 - 7.6.4 Anomaly Correlation 311
 - 7.6.5 Recent Ideas in Nonprobabilistic Field Verification 314
- 7.7 Verification of Ensemble Forecasts 314
 - 7.7.1 Characteristics of a Good Ensemble Forecast 314
 - 7.7.2 The Verification Rank Histogram 316
 - 7.7.3 Recent Ideas in Verification of Ensemble Forecasts 319
- 7.8 Verification Based on Economic Value 321
 - 7.8.1 Optimal Decision Making and the Cost/Loss Ratio Problem 321
 - 7.8.2 The Value Score 324
 - 7.8.3 Connections with Other Verification Approaches 325
- 7.9 Sampling and Inference for Verification Statistics 326
 - 7.9.1 Sampling Characteristics of Contingency Table Statistics 326
 - 7.9.2 ROC Diagram Sampling Characteristics 329
 - 7.9.3 Reliability Diagram Sampling Characteristics 330
 - 7.9.4 Resampling Verification Statistics 332
- 7.10 Exercises 332

CHAPTER 8 Time Series 337

- 8.1 Background 337
 - 8.1.1 Stationarity 337
 - 8.1.2 Time-Series Models 338
 - 8.1.3 Time-Domain vs. Frequency-Domain Approaches 339
- 8.2 Time Domain—I. Discrete Data 339
 - 8.2.1 Markov Chains 339
 - 8.2.2 Two-State, First-Order Markov Chains 340
 - 8.2.3 Test for Independence vs. First-Order Serial Dependence 344
 - 8.2.4 Some Applications of Two-State Markov Chains 346
 - 8.2.5 Multiple-State Markov Chains 348
 - 8.2.6 Higher-Order Markov Chains 349
 - 8.2.7 Deciding among Alternative Orders of Markov Chains 350
- 8.3 Time Domain—II. Continuous Data 352
 - 8.3.1 First-Order Autoregression 352
 - 8.3.2 Higher-Order Autoregressions 357
 - 8.3.3 The AR(2) Model 358
 - 8.3.4 Order Selection Criteria 362
 - 8.3.5 The Variance of a Time Average 363
 - 8.3.6 Autoregressive-Moving Average Models 366
 - 8.3.7 Simulation and Forecasting with Continuous Time-Domain Models 367
- 8.4 Frequency Domain—I. Harmonic Analysis 371
 - 8.4.1 Cosine and Sine Functions 371
 - 8.4.2 Representing a Simple Time Series with a Harmonic Function 372
 - 8.4.3 Estimation of the Amplitude and Phase of a Single Harmonic 375
 - 8.4.4 Higher Harmonics 378
- 8.5 Frequency Domain—II. Spectral Analysis 381
 - 8.5.1 The Harmonic Functions as Uncorrelated Regression Predictors 383
 - 8.5.2 The Periodogram, or Fourier Line Spectrum 383
 - 8.5.3 Computing Spectra 387
 - 8.5.4 Aliasing 388
 - 8.5.5 Theoretical Spectra of Autoregressive Models 390
 - 8.5.6 Sampling Properties of Spectral Estimates 394
- 8.6 Exercises 399

PART III Multivariate Statistics 401

CHAPTER 9 Matrix Algebra and Random Matrices 403

- 9.1 Background to Multivariate Statistics 403
 - 9.1.1 Contrasts between Multivariate and Univariate Statistics 403
 - 9.1.2 Organization of Data and Basic Notation 404
 - 9.1.3 Multivariate Extensions of Common Univariate Statistics 405
- 9.2 Multivariate Distance 406
 - 9.2.1 Euclidean Distance 406
 - 9.2.2 Mahalanobis (Statistical) Distance 407

9.3	Matrix Algebra Review	408
9.3.1	Vectors	409
9.3.2	Matrices	411
9.3.3	Eigenvalues and Eigenvectors of a Square Matrix	420
9.3.4	Square Roots of a Symmetric Matrix	423
9.3.5	Singular-Value Decomposition (SVD)	425
9.4	Random Vectors and Matrices	426
9.4.1	Expectations and Other Extensions of Univariate Concepts	426
9.4.2	Partitioning Vectors and Matrices	427
9.4.3	Linear Combinations	429
9.4.4	Mahalanobis Distance, Revisited	431
9.5	Exercises	432
CHAPTER 10 The Multivariate Normal (MVN) Distribution		435
10.1	Definition of the MVN	435
10.2	Four Handy Properties of the MVN	437
10.3	Assessing Multinormality	440
10.4	Simulation from the Multivariate Normal Distribution	444
10.4.1	Simulating Independent MVN Variates	444
10.4.2	Simulating Multivariate Time Series	445
10.5	Inferences about a Multinormal Mean Vector	448
10.5.1	Multivariate Central Limit Theorem	449
10.5.2	Hotelling's T^2	449
10.5.3	Simultaneous Confidence Statements	456
10.5.4	Interpretation of Multivariate Statistical Significance	459
10.6	Exercises	462
CHAPTER 11 Principal Component (EOF) Analysis		463
11.1	Basics of Principal Component Analysis	463
11.1.1	Definition of PCA	463
11.1.2	PCA Based on the Covariance Matrix vs. the Correlation Matrix	469
11.1.3	The Varied Terminology of PCA	471
11.1.4	Scaling Conventions in PCA	472
11.1.5	Connections to the Multivariate Normal Distribution	473
11.2	Application of PCA to Geophysical Fields	475
11.2.1	PCA for a Single Field	475
11.2.2	Simultaneous PCA for Multiple Fields	477
11.2.3	Scaling Considerations and Equalization of Variance	479
11.2.4	Domain Size Effects: Buell Patterns	480
11.3	Truncation of the Principal Components	481
11.3.1	Why Truncate the Principal Components?	481
11.3.2	Subjective Truncation Criteria	482
11.3.3	Rules Based on the Size of the Last Retained Eigenvalue	484
11.3.4	Rules Based on Hypothesis Testing Ideas	484
11.3.5	Rules Based on Structure in the Retained Principal Components	486

11.4	Sampling Properties of the Eigenvalues and Eigenvectors	486
11.4.1	Asymptotic Sampling Results for Multivariate Normal Data	486
11.4.2	Effective Multiplets	488
11.4.3	The North <i>et al.</i> Rule of Thumb	489
11.4.4	Bootstrap Approximations to the Sampling Distributions	492
11.5	Rotation of the Eigenvectors	492
11.5.1	Why Rotate the Eigenvectors?	492
11.5.2	Rotation Mechanics	493
11.5.3	Sensitivity of Orthogonal Rotation to Initial Eigenvector Scaling	496
11.6	Computational Considerations	499
11.6.1	Direct Extraction of Eigenvalues and Eigenvectors from [S]	499
11.6.2	PCA via SVD	500
11.7	Some Additional Uses of PCA	501
11.7.1	Singular Spectrum Analysis (SSA): Time-Series PCA	501
11.7.2	Principal-Component Regression	504
11.7.3	The Biplot	505
11.8	Exercises	507
CHAPTER 12 Canonical Correlation Analysis (CCA)		509
12.1	Basics of CCA	509
12.1.1	Overview	509
12.1.2	Canonical Variates, Canonical Vectors, and Canonical Correlations	510
12.1.3	Some Additional Properties of CCA	512
12.2	CCA Applied to Fields	517
12.2.1	Translating Canonical Vectors to Maps	517
12.2.2	Combining CCA with PCA	517
12.2.3	Forecasting with CCA	519
12.3	Computational Considerations	522
12.3.1	Calculating CCA through Direct Eigendecomposition	522
12.3.2	Calculating CCA through SVD	524
12.4	Maximum Covariance Analysis	526
12.5	Exercises	528
CHAPTER 13 Discrimination and Classification		529
13.1	Discrimination vs. Classification	529
13.2	Separating Two Populations	530
13.2.1	Equal Covariance Structure: Fisher's Linear Discriminant	530
13.2.2	Fisher's Linear Discriminant for Multivariate Normal Data	534
13.2.3	Minimizing Expected Cost of Misclassification	535
13.2.4	Unequal Covariances: Quadratic Discrimination	537
13.3	Multiple Discriminant Analysis (MDA)	538
13.3.1	Fisher's Procedure for More Than Two Groups	538
13.3.2	Minimizing Expected Cost of Misclassification	541
13.3.3	Probabilistic Classification	542
13.4	Forecasting with Discriminant Analysis	544

- 13.5 Alternatives to Classical Discriminant Analysis 545
 - 13.5.1 Discrimination and Classification Using Logistic Regression 545
 - 13.5.2 Discrimination and Classification Using Kernel Density Estimates 546
- 13.6 Exercises 547

CHAPTER 14 Cluster Analysis 549

- 14.1 Background 549
 - 14.1.1 Cluster Analysis vs. Discriminant Analysis 549
 - 14.1.2 Distance Measures and the Distance Matrix 550
- 14.2 Hierarchical Clustering 551
 - 14.2.1 Agglomerative Methods Using the Distance Matrix 551
 - 14.2.2 Ward's Minimum Variance Method 552
 - 14.2.3 The Dendrogram, or Tree Diagram 553
 - 14.2.4 How Many Clusters? 554
 - 14.2.5 Divisive Methods 559
- 14.3 Nonhierarchical Clustering 559
 - 14.3.1 The K-Means Method 559
 - 14.3.2 Nucleated Agglomerative Clustering 560
 - 14.3.3 Clustering Using Mixture Distributions 561
- 14.4 Exercises 561

APPENDIX A Example Data Sets 565

- Table A.1. Daily precipitation and temperature data for Ithaca and Canandaigua, New York, for January 1987 565
- Table A.2. January precipitation data for Ithaca, New York, 1933–1982 566
- Table A.3. June climate data for Guayaquil, Ecuador, 1951–1970 566

APPENDIX B Probability Tables 569

- Table B.1. Cumulative Probabilities for the Standard Gaussian Distribution 569
- Table B.2. Quantiles of the Standard Gamma Distribution 571
- Table B.3. Right-tail quantiles of the Chi-square distribution 576

APPENDIX C Answers to Exercises 579

References 587

Index 611